EXTRACTING MEANINGFUL STATISTICS FOR THE CHARACTERIZATION AND CLASSIFICATION OF BIOLOGICAL, MEDICAL, AND FINANCIAL DATA

A Thesis
Presented to
The Academic Faculty

by

Tonya M. Woods

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2015

EXTRACTING MEANINGFUL STATISTICS FOR THE CHARACTERIZATION AND CLASSIFICATION OF BIOLOGICAL, MEDICAL, AND FINANCIAL DATA

Approved by:

Professor Brani Vidakovic, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Kamran Paynabar
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Mirjana Milosevic-Brockett
School of Biology
Georgia Institute of Technology

Dr. Scott Nickleach
Statistical Consultant
Equifax INC

Date Approved: May 1, 2015

Dedicated to my Papa Don, Class of 1958 . . .

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Brani Vidakovic. Ever since my first day in his biostatistics course, he has been completely supportive of my work and my aspirations. Thank you for always greeting me with a smile and a handshake. I would also like to thank Dr. Scott Nickleach, Dr. William Auffermann, MD, Mary Newell, MD, Dr. Thanawadee Preeprem, Dr. Kichun Lee, Dr. Woojin Chang, and Minkyoung Kang for their contributions to this work. In addition, I want to express my gratitude for my thesis committee members, Dr. Yajun Mei, Dr. Kamran Paynabar, Dr. Mirjana Milosevic-Brockett, and Dr. Scott Nickleach, for their careful review of this work and their insightful comments.

Thank you to my amazing husband, Paul Woods, who has both financially and emotionally supported me through my entire graduate school experience. Thank you for teaching me the value of a nice graphic, particularly one with the proper use of color. I'm sorry to have kept you up late at night with my worries of not passing the comps or of not ever finishing my dissertation.

I would also like to thank my parents, Sam Roberts (class of 1983) and Marcia Roberts (class of 1986), who have diligently read every paper I have written and have always provided meaningful, thought-provoking feedback. Thank you for always believing in my capabilities and continuing to support my academic growth. And lastly, thank you for reminding me to take breaks and eat meals.

Thank you to all of my family members, who have been and continue to be major influences in my life. Nanny and Papa Bill: I am so lucky to have you nearby to nurture me and keep me well fed! Debbie and Jerry: Thank you for encouraging me to get my doctorate and always celebrating my successes. Mimi: Thank you for our

wonderful phone conversations. I always hang up feeling more positive and assured than before. Pam (class of 1988): Thank you for your perpetual confidence in me. Finally, I would like to thank my Papa Don, who graduated in the class of 1958 with a degree in Industrial Engineering from Georgia Tech. Thank you for starting the fine tradition in our family of attending such a prestigious university. I hope the work I have done while at Georgia Tech would make you proud.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . . **iv**

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **x**

SUMMARY . . . . . **xiii**

I BACKGROUND ON SCALING AND WAVELETS **1**

1.1 Self-Similar Processes 1

1.1.1 Brownian Motion (Bm) and Fractional Brownian Motion (fBm) 2

1.1.2 Applications to Real-World Processes 3

1.2 The Traditional Wavelet Transform 7

1.2.1 The Discrete Complex Wavelet Transform 8

1.3 The Traditional 2-D Wavelet Transform 11

1.4 The Scale-Mixing 2-D Wavelet Transform 16

1.4.1 Definition of Scale-Mixing Wavelet Spectra 19

II CHARACTERIZING EXONS AND INTRONS BY REGULARITY OF NUCLEOTIDE STRINGS **22**

2.1 Introduction 22

2.1.1 Exons and Introns in Eukaryotic DNA 22

2.1.2 Previous Work on Translating DNA Nucleotides to Numbers 23

2.2 From ACGT to Numbers 25

2.2.1 Translating to Matrices via Assignment of Unit Vectors . . . 25

2.2.2 Equivalence Classes 28

2.2.3 Invariant Translation Procedure 29

2.2.4 Cumulative Evolutionary Slope 30

2.3 Application 30

2.3.1 Data 31

2.3.2 Comparing Regularity of Honeybee and Simulated DNA . . . 32

2.3.3 Comparing Regularity of Exons and Introns 35
2.4 Discussion 39
III WAVELET-BASED SCALING INDICES FOR CANCER DIAGNOSTICS 41
3.1 Ovarian Cancer Diagnostics 41
3.1.1 Introduction 41
3.1.2 Data 43
3.1.3 Time-varying Slope 45
3.1.4 Phase 49
3.1.5 Combining Classification Procedures 50
3.1.6 Classification Results 51
3.2 Breast Cancer Diagnostics 52
3.2.1 Introduction 52
3.2.2 Data 54
3.2.3 The Scale-Mixing Transform and Spectral Slope 55
3.2.4 Asymmetry Statistics 56
3.2.5 Comparing Descriptors for Cases and Controls 61
3.2.6 Classification Results 62
3.2.7 Conclusion 63
3.3 Lung Cancer Diagnostics 64
3.3.1 Introduction 64
3.3.2 Data 65
3.3.3 The Scale-Mixing Transform, Spectral Slope, and Asymmetry Statistic 66
3.3.4 Classification Results 66
IV ASSESSING THE IMPACT OF SOCIAL MEDIA DATA IN COMMERCIAL CREDIT MODELING VIA RANDOM FORESTS . 69
4.1 Introduction 69
4.1.1 Typical Scoring Methodology 70

4.1.2 The Emergence of Social Media Data in Decision-Making . . 70
4.2 Background and Considerations 72
4.2.1 Data Regulations and Restrictions 74
4.2.2 Model Requirements and Methodology 75
4.3 Data Collection 75
4.3.1 Collecting Online Hotel Reviews 76
4.3.2 Hotel Attribute Creation 77
4.3.3 Merging to Bureau Data 79
4.3.4 Dependent Variables 79
4.4 Modeling Methodology and Results 80
4.4.1 Random Forest Overview 80
4.4.2 Our Modeling Approach 83
4.4.3 Tuning Parameters 84
4.4.4 Results 86
4.5 Discussion and Conclusions 92
APPENDIX A — EXPRESSION 20 PROOF 94
BIBLIOGRAPHY 96
VITA 102

LIST OF TABLES

1 Three equivalence classes of assignments of nucleotides to unit vectors that lead to three different slopes, s_1, s_2 and s_3. 29

2 Classification results for the cut-off slope value, $s^* = -1.735$, which maximizes the Youden Index. 38

3 Classification results 52

4 Comparison of the asymmetry statistics for images with strong vertical and horizontal features by dyadic level pairing. The systematic differences in asymmetry statistics may be seen for the coarser level pairings (2 and 3, 3 and 4, 4 and 5). The finer level pairings (5 and 6, 6 and 7) are not capable of "seeing" the images' directional differences. 59

5 Two-way nested ANOVA on spectral slopes 61

6 Two-way nested ANOVA on t statistics 62

7 Two-way nested ANOVA on fc statistics 62

8 SVM classification results. The best results are achieved using the linear kernel and including both spectral slopes and fold change asymmetry statistics in the classification. 63

9 SVM classification results - lung CXRs 67

10 Bad capture rates for the social media data only, traditional bureau data only, and combined default models. 88

11 Cumulative bad dollar capture for the social media data only, traditional bureau data only, and combined default models. 88

12 Diagnostics for the default models 89

13 Good capture rates for the social media data only, traditional bureau data only, and combined activity models. 90

14 Cumulative good dollar capture for the social media data only, traditional bureau data only, and combined activity models. 90

15 Diagnostics for the activity models 91

16 Top five most influential attributes in the social media only and bureau only default models. 91

17 Top five most influential attributes in the social media only and bureau only activity models. 92

LIST OF FIGURES

1 Simulated paths of fractional Brownian motion, (a) $H = 1/4$, (b) $H = 1/2$, and (c) $H = 3/4$. 3

2 Nile yearly minimal level data (left) and its wavelet log spectra (right) 5

3 Coke stock market prices (left), its scaling behavior in the Fourier domain (center) and in the wavelet domain (right). 5

4 (a) Exchange rates HKD per US$ (left), its scaling behavior in the Fourier domain (center) and in the wavelet domain (right) 6

5 Gait timing for Slow, Normal and Fast Walk (left), scaling behavior in the Fourier domain (center) and in the wavelet domain (right). 7

6 Doppler signal: Real and imaginary parts of the complex wavelet transform using the complex filter Symmetric Daubechies Wavelets (SDW) 6. 9

7 Top left: Original Lenna image; Top right: Original Barbara image; Bottom left: Image with Lenna modulus but Barbara phase; Bottom right: Image with Barbara modulus but Lenna phase 10

8 Tessellations for some 2-D wavelet transform of depth 4. 13

9 Steps of a two-level discrete wavelet transform of an image: wavelet transform on all rows (left); one-level decomposition (center); two-level decomposition (right). 13

10 One step in the wavelet transformation of an image. The standard image processing template image Lenna is used. 15

11 Image of a box with a cross within (left), and its 2D wavelet transform using COIFLET2 basis and $L = 3$ levels of resolutions (right). 16

12 Tessellations for some 2-D scale-mixing wavelet transform of depth 4 (left), and 2-D scale-mixing wavelet transform of the box with cross image (right). 18

13 Three detail-space hierarchies generating the scale-mixing 2-D transform, where (j_1, j_2) is indexed as $(j, j + s)$, $s \in \mathbb{Z}$. Circles correspond to $s = 0$, triangles to $s = 1$, and squares to $s = -1$; 20

14 Illustration of submatrices Z_1 and Z_2, the levels of details forming the hierarchy for defining the log-spectral slope, in the original (top) and modified (bottom) procedures. 28

15 Cumulative evolutionary slope calculation. Overlapping sequences of DNA nucleotides are represented as matrices, the scale-mixing wavelet transformation is applied to these matrices, log average energies are computed for the shown detail levels, and slopes are calculated. . . . 31

16 Global slope for the honeybee DNA (red line at -1.7825) sequence and empirical distribution of slopes for 10,000 simulated random DNA-like sequences of length 2^9. 33

17 (Left) Honeybee cumulative evolutionary slope for gene "LOC100577807"; (Right) Cumulative evolutionary slope for a random DNA-like sequence; In both cases window size was 2^5 34

18 An illustration of detail levels of Z used in the slope calculation for window size 32, honeybee DNA. 34

19 Honeybee cumulative evolutionary slope for gene "LOC408625" on the first chromosome. Solid red line: Average slope for a coding sequence (exons), solid green line: Average slope for a noncoding sequence (introns), solid blue line: Average slope for a sequence with a mixture of coding and noncoding, dotted red line: Division between type (exons, introns, combination) of region. 35

20 Honeybee cumulative evolutionary slope for genes "Ard1" and "Dat" on the first chromosome. Solid red line: Average slope for a coding sequence (exons), solid green line: Average slope for a noncoding sequence (introns), solid blue line: Average slope for a sequence with a mixture of coding and noncoding, dotted red line: Division between type (exons, introns, combination) of region. 37

21 Kernel density estimates for cumulative evolutionary slopes of exons (red) and introns (green) 38

22 Illustration of LC-MS. 44

23 Illustration of the mass spectrometry data. 45

24 Examples of ion feature signals for cases. 45

25 Examples of ion feature signals for controls. 46

26 Time-varying log-spectral slopes of ion feature signals for cases (blue) and controls (green). 47

27 Regions of time-varying log-spectral slopes used to create classification inputs. 48

28 ROC curve: True positive rate against false positive rate as the classification threshold is varied. 50

29 Combining classification procedures: serial. 51

30 Combining classification procedures: parallel. 51

31 Mammogram images, with cancer (left) and without cancer (right), split into five sub-images each. 54

32 Log energy spectra for a single region of a case (left) and the log energy spectra for the corresponding region of a control (right) 55

33 Examples of images with low (left), moderate (middle), and high (right) regularity. 56

34 Illustration of wavelet coefficients contributing to the asymmetry statistics. Red lines connect the regions of the wavelet-transformed image used in the calculations. 57

35 Image with strong vertical features (left) and image with strong horizontal features (right). 58

36 Chest radiograph (left) and the portion analyzed for deviation from isotropy (right) 59

37 Empirical bootstrap distributions (histograms) for the t statistic at the three coarsest levels and where the chest radiograph subimage's asymmetry statistics fall in the distributions (red vertical lines). This image has asymmetry statistics falling in the left tails of the distributions, indicating high horizontal directionality (t statistic ASLs (coarse to fine): 0.061, 0.001, 0). 60

38 Empirical bootstrap distributions (histograms) for the fc statistic at the three coarsest levels and where the chest radiograph subimage's asymmetry statistics fall in the distributions (red vertical lines). This image has asymmetry statistics falling in the left tails of the distributions, indicating high horizontal directionality (fc statistic ASLs (coarse to fine): 0.051, 0, 0). 60

39 Lung CXR image (left) and mask showing pulmonary nodule location (right) 65

40 ROI (left) and ROC (right) 66

41 Log energy spectra for ROI (left) and ROC (right) 67

42 Binary decision tree 81

43 A comparison of the models' performances for the default dependent variable (top) and the activity dependent variable (bottom). Note that for both sets of models, the bureau only and combined curves are overlapping. 87

SUMMARY

This thesis is focused on extracting meaningful statistics for the characterization and classification of biological, medical, and financial data. It comprises three main topics:

1. Characterizing exons and introns by regularity of nucleotide strings

In this work, we outline a methodology for representing sequences of DNA nucleotides as numeric matrices in order to analytically investigate important structural characteristics of DNA. This methodology involves assigning unit vectors to nucleotides, placing the vectors into columns of a matrix, and accumulating across the rows of this matrix. Transcribing the DNA in this way allows us to compute the 2-D wavelet transformation and assess regularity characteristics of the sequence via the slope of the wavelet spectra. Based on the existence of equivalence classes, we adjust the methodology so that the log-spectral slope does not depend on the assignment of unit vectors. In addition to computing a global slope measure for a sequence, we can apply our methodology for overlapping sections of nucleotides to obtain an evolutionary slope. To illustrate our methodology, we analyze 376 gene sequences from the first chromosome of the honeybee. First, we simulate "random DNA" using actual proportions of the nucleotides from the honeybee DNA and find that the actual DNA sequence is more regular than simulated DNA sequences. The second analysis involves calculation of the cumulative evolutionary slope for each gene. We label each window of DNA as exons (coding regions), introns (noncoding regions), or a combination of the two and average the cumulative evolutionary slopes for each of the three categories. For the genes analyzed, we find that introns are significantly

more regular (lead to more negative slopes) than exons, which agrees with the results from the literature where regularity is measured on "DNA walks."

2. Wavelet-based scaling indices for cancer diagnostics

There were nearly half of a million new cases of ovarian, breast, and lung cancer in the United States last year. Breast and lung cancer have highest prevalence, while ovarian cancer has the lowest survival rate of the three. Early detection is critical for all of these diseases, but substantial obstacles to early detection exist in each case.

Ovarian cancer, the "silent killer," is most challenging to detect in early stages since most women do not experience any symptoms early on. In addition, there is no official screening procedure for ovarian cancer. In this work, we investigate the use of metabolic data for detecting ovarian cancer. Through wavelet-based scaling, we produce a classification procedure using time-varying slopes of the metabolic series with approximately 65% accuracy (sensitivity=60%, specificity=70%). In addition, we combine this procedure in parallel with one based on phase information to achieve a procedure with comparable accuracy, but very high sensitivity (approximately 89%). This procedure may be useful for identifying which individuals should be more closely monitored for the disease.

Mammography is used to screen for breast cancer. However, the radiological interpretation of mammogram images is complicated by the heterogeneous nature of normal breast tissue and the fact that cancers are often of the same radiographic density as normal tissue. In this work, we use wavelets to quantify spectral slopes of BC cases and controls and demonstrate their value in classifying images. In addition, we propose asymmetry statistics to be used in forming additional features which improve the classification result. For the best classification procedure, we achieve approximately 77% accuracy (sensitivity=73%, specificity=84%) in classifying mammograms with and without cancer.

Pulmonary nodules, indicative of lung cancer, may be identified on chest X-rays (CXRs). However, even with considerable training, radiologists often overlook nodules. An estimated 75% of perihilar and 90% of peripheral nodules that are identifiable on lung CXRs are missed at the time of initial interpretation. In this work, we investigate the spectral slopes and asymmetry statistics of regions of interest (ROIs) and regions of control (ROCs) and assess their value in classifying images. For the best classification procedure, we achieve approximately 59% accuracy (sensitivity=65%, specificity=52%) in classifying lung cancer cases and controls. This lower accuracy rate, in comparison with the breast cancer classification, may be due to an insufficient sample size or the heterogeneity of anatomic structures in the images.

Computer-aided detection (CAD) algorithms for detecting lung and breast cancer often focus on select features in an image and make a priori assumptions about the nature of a nodule or a mass. In contrast, our approach to analyzing breast and lung images captures information contained in the background tissue of images as well as information about specific features and makes no such a priori assumptions about a nodule or mass.

3. Assessing the impact of social media data in commercial credit modeling via random forests

We investigate the value of social media data in building commercial default and activity credit models. We use random forest modeling, which has been shown in many instances to achieve better predictive accuracy than logistic regression in modeling credit data. This result is of interest, as some entities are beginning to build credit scores based on this type of publicly available online data alone. Our default dependent variable is based on 90 day or more past due account delinquency within a six month period, and our activity dependent variable is based on whether a company is likely to open a new account in the near future. We build the first set of models

via random forests using a large set of public record, firmographic, and non-financial data from a major financial information services provider and tune the parameters to maximize predictive power. The second set of models is constructed in the same way, but includes only social media data derived from online reviews. We then build a third set of models using both bureau and social media data. Our work has shown that the addition of social media data does not provide any improvement in model accuracy (measured by KS, lift in the Lorenz curve on the testing set) over the bureau only models. However, the social media data on its own does have some limited predictive power.

CHAPTER I

BACKGROUND ON SCALING AND WAVELETS

1.1 Self-Similar Processes

We assume that all processes discussed are real valued and defined on the same parameter space. Two processes $X(t)$ and $Y(t)$, *equal in all finite dimensional distributions*, will be denoted as $X(t) \stackrel{d}{=} Y(t)$. This means that for any selection of "times" $0 \leq t_1 < t_2 < \ldots t_k < \infty$, random vectors $(X(\omega, t_1), \ldots, X(\omega, t_k))$ and $(Y(\omega, t_1), \ldots, Y(\omega, t_k))$ have the same distribution. Informally, processes equal-in-distribution are statistically indistinguishable.

Random process $X(t)$ is called stochastically continuous at t_0 if

$$\lim_{h \to 0} P(|X(t_0 + h) - X(t_0)| > \epsilon) = 0$$

for any fixed $\epsilon > 0$.

Also, we consider processes not to be trivial. [1]

A random process $X(t), t > 0$ is called *self-similar* if for any $a > 0$, there exists $b > 0$ such that

$$X(at) \stackrel{d}{=} bX(t). \tag{1}$$

Lamperti (1972) proved the result:

If random process $X(t), t \geq 0$ is nontrivial, stochastically continuous at 0, and self-similar, then there exists unique $H \geq 0$ such that $b = a^H$. If $X(0) = 0, a.s.$ then $H > 0$.

Therefore, a standard definition of self-similar processes is as follows: Process $X(t),\ t \geq 0$ is self-similar, with self-similarity index H (H-ss) if and only if there exists

[1] Process $X(t)$ is trivial if the distribution of random variable $X(\omega, t),\ t$ fixed is a point mass measure. For example, $X(t) = \text{const}$ or $X(t) = \sin(t)$ would be examples of trivial processes.

an $H > 0$ such that for any $a > 0$, $X(at) \stackrel{d}{=} a^H X(t)$. Also see, for example, Beran, 1994; Samorodnitsky and Taqqu, 1994; Oppenheim and Taqqu, 2003; Embrechts and Maejima, 2002). Uniqueness of H is not obvious from this definition, although, H is unique by Lamperti's theorem. Also, from Definition 1.1 it follows that $X(0) = 0$.

1.1.1 Brownian Motion (Bm) and Fractional Brownian Motion (fBm)

Brownian motion $B(t)$ is standardly defined as a random process satisfying the following four requirements:

(i) $B(0) = 0$,

(ii) For any choice n and $0 \leq t_1 < t_2 < \cdots < t_n$, the increments $B(t_2) - B(t_1), \ldots, B(t_n) - B(t_{n-1})$ are independent and stationary;

(iii) For fixed t, $B(t)$ is the Gaussian random variable with zero mean and variance t, and

(iv) $B(t)$ is a continuous function of t, a.s.

It is straightforward to check that the Brownian motion is an 1/2-ss process, for $B(t) = a^{-1/2}B(at)$ conforms to properties (i)-(iv).

A zero mean Gaussian process $B_H(t)$ is called fractional Brownian motion with Hurst exponent H, if

$$EB_H(t)B_H(s) = \frac{E|B_H(1)|^2}{2}\left[|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right],$$

where $E|B_H(1)|^2 = \frac{\Gamma(2-2H)\cos(\pi H)}{\pi H(1-2H)}$.

The process $B_H(t)$ is unique, in the sense that the class of all fractional Brownian motions with exponent H coincides with the class of all Gaussian H-ss processes. However, a Gaussian process is H-ss with independent increments, if and only if it $H = 1/2$, i.e., if it is a Brownian motion.

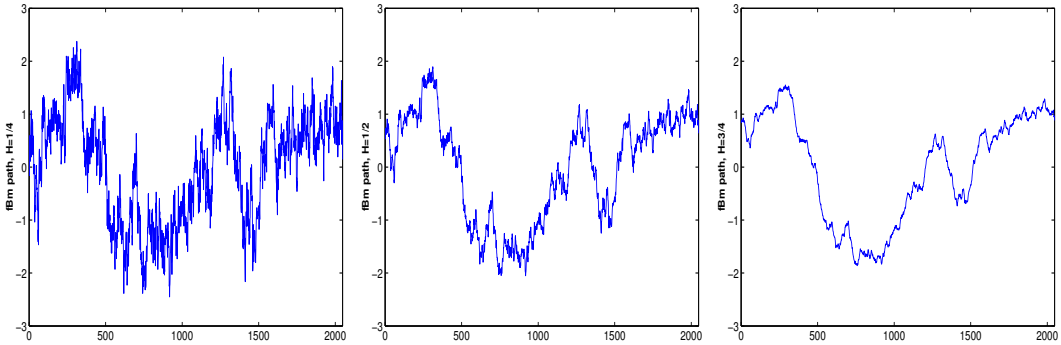


Figure 1: Simulated paths of fractional Brownian motion, (a) $H = 1/4$, (b) $H = 1/2$, and (c) $H = 3/4$.

Sample paths of Brownian motion and fractional Brownian motion behave similarly. Figure 1 shows simulated paths for different values of H. They are continuous almost surely for all $H \in (0, 1)$ and nowhere differentiable. For small H (say, $H < 0.5$) the sample paths are quite irregular and *space-filling.*

1.1.2 Applications to Real-World Processes

Theoretical self-similar processes (such as fractional Brownian motion) are becoming fundamental in modeling of wide-range of real-world phenomena in engineering, physics, medicine, biology, engineering, art, economics, astronomy, chemistry, etc. Time series can be explored in two complementary domains: time and scale/frequency domain. It is mostly the second domain that reveals the scaling and self-similarity properties of time series.

1.1.2.1 It Started with Hurst and Nile Data

British hydrologist Harold Edwin Hurst spent 62 years in Egypt and mostly worked on design and construction of reservoirs along the Nile River. By inspecting historical data on the Nile River flows, Hurst discovered phenomenon (now called Hurst effect).

Hurst was trying to find an optimal reservoir capacity R such that it can accept the river flow in N units of time, $X_1, X_2, \ldots X_N$, and have a constant withdrawal of

$\bar{X}$ per unit time. The optimal volume of the reservoir was given by the so called adjusted range,

$$R = \max_{1 \le k \le N}(X_1 + \cdots + X_k - k\bar{X}) - \min_{1 \le k \le N}(X_1 + \cdots + X_k - k\bar{X}) \tag{2}$$

Since records for waterflow rarely exceeded 100 years Hurst inspected other geophysical data and, in order to compare them, he standardized their adjusted ranges R, with sample standard deviation

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}\ , \tag{3}$$

to obtain dimensionless ratio R/S - rescaled and adjusted range.

On basis of more that 800 records, he found (Hurst, 1951) that quantity R/S scales as N^H, for ranging from 0.46 to 0.93, with mean 0.73 and standard deviation of 0.09.

This result was in contrast to the fact that for independent normal random variables H is 1/2 in the limit. Feller (1951) proved that the limit is 1/2 for independent identically distributed random variables with finite second moment. It was believed that strong Markovian dependence was responsible for deviations from $H = 1/2$ until Barnard (1956) proved that limit $H = 1/2$ holds for the Markovian dependence case. It was the later work of Mandelbrot (1975), Mandelbrot and J. W. Van Ness (1968), and Mandelbrot and Wallis (1969) who associated the Hurst (or Joseph) phenomenon with the presence of long-memory of a time series.

Figure 2 (left) shows n=512 consecutive yearly measurements from the famous Nile River Data set for the years 62-1281 A.D. Figure 2 (right) shows its wavelet spectra, demonstrating the scaling law.

1.1.2.2 Economic Time Series

Many economic time series, such as stock market prices, exchange rates and asset returns exhibit scaling laws and long range dependence (LRD). This is in empirical

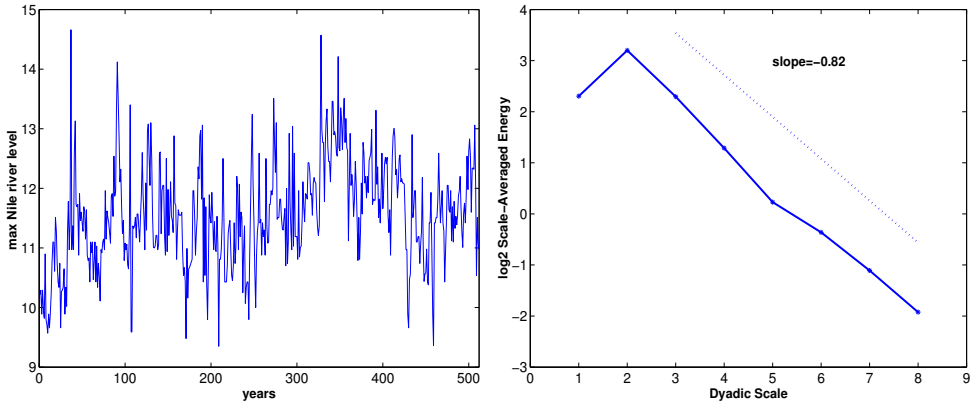


Figure 2: Nile yearly minimal level data (left) and its wavelet log spectra (right)

contradiction to several economic theories (random walk theory for stock market, perfect markets, etc) and gave rise to several theories and models describing the scaling and LRD (such as ARFIMA, fGn, fBm, GARCH, etc.).

We present two data sets: Coca Cola stock market prices (Figure 3 (left)) and rates of exchange between Hong Kong Dollar (HKD) and USDollar (USD) (Figure 4 (left)), as reported by the ONADA Company between March 24, 1995 and November 1, 2000.

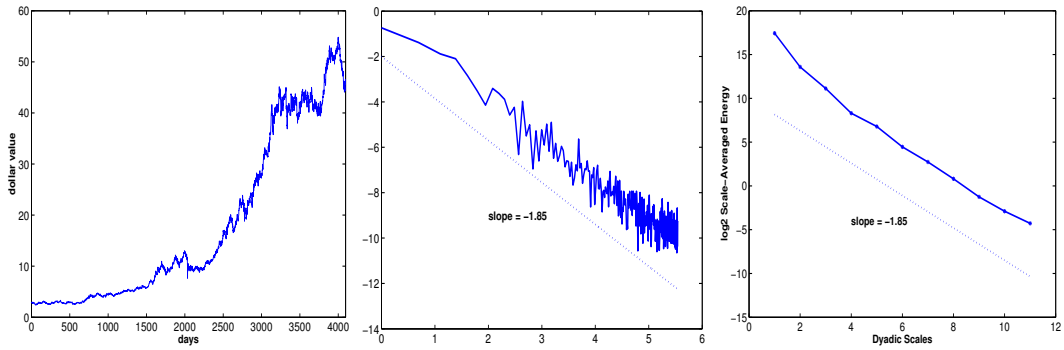


Figure 3: Coke stock market prices (left), its scaling behavior in the Fourier domain (center) and in the wavelet domain (right).

The data relative to the rates of exchange between (HKD) and USDollar (USD) are from The Bank of Korea Economic Statistics System

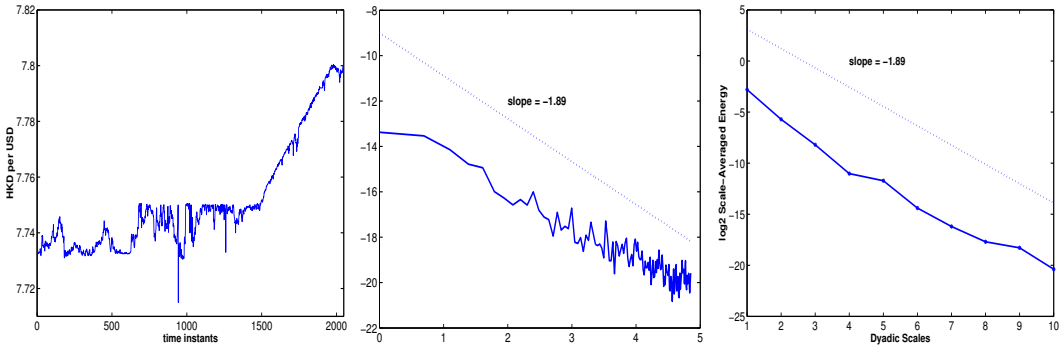


Figure 4: (a) Exchange rates HKD per US$ (left), its scaling behavior in the Fourier domain (center) and in the wavelet domain (right)

(http://ecos.bok.or.kr/EIndex_en.html)

Corresponding Fourier and wavelet spectra are shown in the center and right panels of Figures 3 and 4.

1.1.2.3 Gait Data

Scaling laws were recently detected in the apparently "noisy" variations in the stride interval (duration of the gait cycle) of human walking. Dynamic analysis of these step-to-step fluctuations revealed a self-similar pattern: Fluctuations at one time scale are statistically similar to those at multiple other time scales, at least over hundreds of steps, while healthy subjects walk at their normal rate. The experimental data consist of measurements for a healthy subject who walked for 1 hour at his usual, slow and fast paces. The stride interval fluctuations exhibited long-range correlations with power-law decay for up to a thousand strides at all three walking rates.

It is curious that during metronomically-paced walking, these long-range correlations disappeared; variations in the stride interval were anti-correlated. Experiments confirm that scaling behavior of spontaneous stride interval are normally quite robust and intrinsic to the locomotor system. Furthermore, this fractal property of neural output may be related to the higher nervous centers responsible for control of walking rhythm.

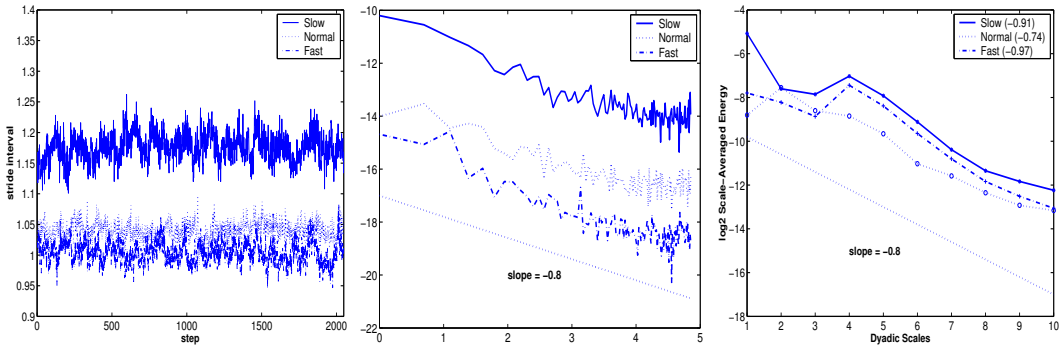


Figure 5: Gait timing for Slow, Normal and Fast Walk (left), scaling behavior in the Fourier domain (center) and in the wavelet domain (right).

Participants in this experiment had no history of any neuromuscular, respiratory or cardiovascular disorders, and were taking no medications. Mean age was 21.7 years (range: 18-29 years). Height was 1.77 $\pm$ 0.08 meters (mean +/- S.D.) and weight was 71.8 $\pm$ 10.7 kg. Subjects walked continuously on level ground around an obstacle free, long (either 225 or 400 meters), approximately oval path and the stride interval was measured using ultra-thin, force sensitive switches taped inside one shoe. Figure 5 shows 2048 data points for one subject. Slow and fast stride intervals have slopes of -0.91 and -0.97 respectively, and stride intervals for normal walk show scaling with -0.74 slope.

1.2 The Traditional Wavelet Transform

The discrete wavelet transform (DWT) expresses a real time series $X(t)$ in terms of shifted and dilated versions of a wavelet (or *mother*) function $\psi(t)$ and shifted versions of a scaling (or *father*) function $\phi(t)$. For specific choices of the scaling functions and wavelets, an orthonormal basis can be formed from the atoms

$$\begin{aligned} \psi_{j,k}(t) &= 2^{j/2}\,\psi(2^j t - k) \\ \phi_{j,k}(t) &= 2^{j/2}\,\phi(2^j t - k), \quad j,k \in \mathbb{Z}. \end{aligned}$$

Then, $X(t)$ can be represented by wavelets as

$$X(t) = \sum_k c_{J_0,k}\phi_{J_0,k}(t) + \sum_{j=J_0}^{\infty}\sum_k d_{j,k}\psi_{j,k}(t),$$

where

$$d_{j,k} = \int X(t)\ \psi_{j,k}(t)\ dt, \qquad c_{j,k} = \int X(t)\ \phi_{j,k}(t)\ dt, \tag{4}$$

are detail and scaling coefficients respectively. Here, J_0 indicates the coarsest scale or lowest resolution level of the transform, and larger values of j correspond to higher resolutions. For a detailed introduction to wavelet theory, the reader is referred to Daubechies (1992) or Mallat (1997).

1.2.1 The Discrete Complex Wavelet Transform

The discrete complex wavelet transform (DCWT) can be considered a complex-valued extension to the standard discrete wavelet transform (DWT) which uses complex-valued filtering (analytic filter) for decomposing the real/complex signals into real and imaginary parts in transform domain. Complex wavelet coefficients can be computed using Mallat's algorithm,

$$c_{j-1,l} = \sum_k \overline{h}_{k-2l}c_{j,k}. \tag{5}$$

and

$$d_{j-1,l} = \sum_k \overline{g}_{k-2l}c_{j,k} \tag{6}$$

Conversely, the reconstruction is given by

$$c_{j,k} = \sum_l c_{j-1,l}h_{k-2l} + \sum_l d_{j-1,l}g_{k-2l}. \tag{7}$$

Unlike the real case, the complex wavelet representation provides a redundant description of the signal. Moreover, the real and imaginary coefficients are used to

compute amplitude and phase information. Let's call

$$\psi^c(x) = \psi^r(x) + i\psi^i(x)$$

the complex wavelet. Projecting the signal onto $2^{\frac{j}{2}}\psi^c(2^j x - k)$, we obtain the complex wavelet coefficient

$$d^c_{j,k} = d^r_{j,k} + id^i_{j,k}$$

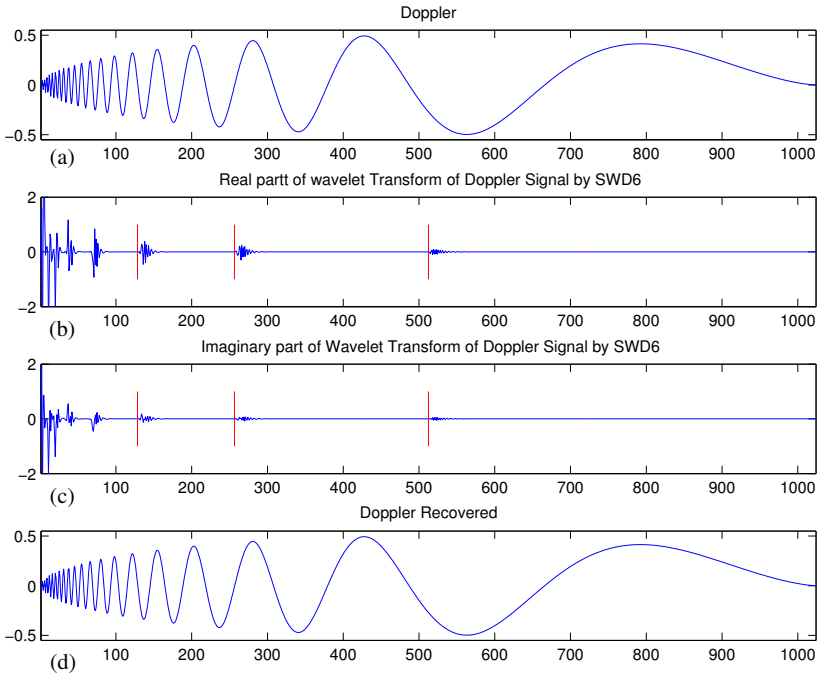


Figure 6: Doppler signal: Real and imaginary parts of the complex wavelet transform using the complex filter Symmetric Daubechies Wavelets (SDW) 6.

with magnitude (or modulus)

$$|d^c_{j,k}| = \sqrt{(d^r_{j,k})^2 + (d^i_{j,k})^2}$$

and phase

$$\angle d^c_{j,k} = \arctan\left(\frac{d^i_{j,k}}{d^r_{j,k}}\right).$$

when $|d^c_{j,k}| > 0$.

The DCWT enables new coherent multiscale signal processing algorithms that exploit the complex magnitude and phase. Figure 6 shows an application of functions to the Doppler signal using the complex filter Symmetric Daubechies Wavelets (SWD) 6.

Complex wavelet coefficients can also be obtained by using the matrix multiplication $\mathbf{d} = W_j \cdot \mathbf{y}$ where W_j is a complex wavelet matrix based on the filters H_k and G_k. The high-pass filter is obtained as $g_k = (-1)^k \bar{h}_{1-k}$.

The DCWT may be applied in the 2-dimensional case as well (i.e. to images; more details on generalizing the wavelet transform to the 2-dimensional case are given in the next section). In fact, complex wavelet-based approaches have recently been successfully applied to image denoising, restoration, and enhancement and have in some cases outperformed methods using the regular DWT (Hostalkova and Prochazka, 2007).

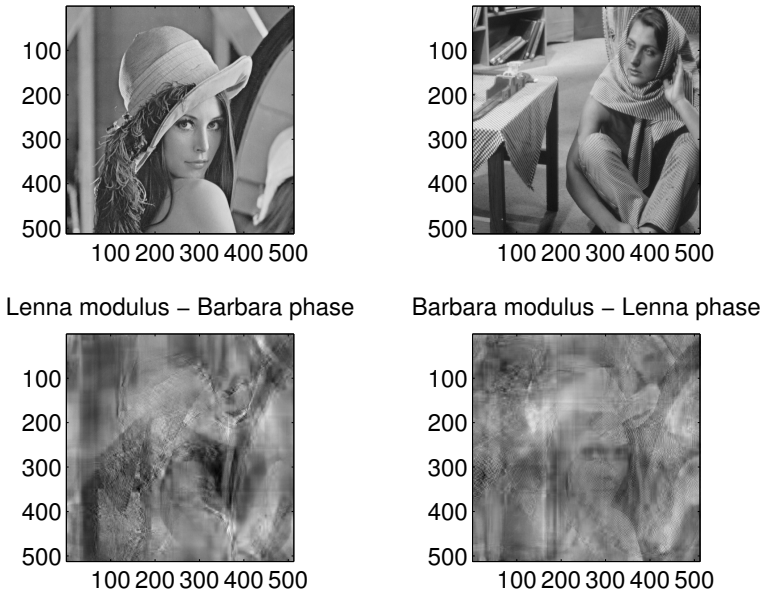


Figure 7: Top left: Original Lenna image; Top right: Original Barbara image; Bottom left: Image with Lenna modulus but Barbara phase; Bottom right: Image with Barbara modulus but Lenna phase

Although it is sometimes suggested that phase information is of less significance than magnitude, Figure 7 demonstrates otherwise. Figure 7 (top) shows two standard image processing template images, Lenna and Barbara. Figure 7 (bottom) shows the result when the magnitudes and phases are intermixed. Interestingly, the image with phase information input into the mixed image is visually prominent in each case.

1.3 The Traditional 2-D Wavelet Transform

Many time series arising in practical applications are multidimensional. Examples include medical imaging taken at equispaced time instants, where the spatial structure is on a regular grid of pixels. The wavelet transform is readily generalized to the multidimensional case; the 2-D wavelet basis functions for traditional 2-D wavelet transform are constructed via translations and dilations in a product of univariate wavelet and scaling functions

$$\begin{aligned} \phi(t_1,t_2) &= \phi(t_1)\phi(t_2) \\ \psi^h(t_1,t_2) &= \phi(t_1)\psi(t_2) \\ \psi^v(t_1,t_2) &= \psi(t_1)\phi(t_2) \\ \psi^d(t_1,t_2) &= \psi(t_1)\psi(t_2). \end{aligned} \tag{8}$$

The symbols h, v, d in (8) stand for horizontal, vertical and diagonal directions, respectively, since the atoms capture image features in the corresponding directions.

Consider the wavelet atoms

$$\phi_{j,\mathbf{k}}(\mathbf{t}) = 2^j\ \phi(2^j t_1 - k_1, 2^j t_2 - k_2) \tag{9}$$

$$\psi^i_{j,\mathbf{k}}(\mathbf{t}) = 2^j\ \psi^i(2^j t_1 - k_1, 2^j t_2 - k_2), \tag{10}$$

for $i = h, v, d, \quad j \in \mathbb{Z}$, $\mathbf{t} = (t_1, t_2) \in \mathbb{R}^2$, and $\mathbf{k} = (k_1, k_2) \in \mathbb{Z}^2$. Then, any function $X \in \mathcal{L}_2(\mathbb{R}^2)$ can be represented as

$$X(\mathbf{t}) = \sum_{\mathbf{k}} c_{J_0\mathbf{k}}\phi_{J_0,\mathbf{k}}(\mathbf{t}) + \sum_{j\geq J_0}\sum_{\mathbf{k}}\sum_{i} d^i_{j,\mathbf{k}}\psi^i_{j,\mathbf{k}}(\mathbf{t}), \tag{11}$$

where the wavelet coefficients are given by

$$d^i_{j,\mathbf{k}} = 2^j \int X(\mathbf{t})\ \psi^i(2^j\mathbf{t} - \mathbf{k})\ d\mathbf{t}.$$

It is well known that discrete wavelet transform, similar to discrete Fourier transform, can be achieved by matrix multiplication – since both transforms are linear. We briefly describe the construction of the wavelet matrix in the one-dimensional case, and direct the reader to Vidakovic (1999) (pp 115-116, 153-159) for details and the higher dimensional case.

Let the length of the univariate time series $\mathbf{y}$ be 2^J and $\mathbf{h} = \{h_s,\ s \in \mathbb{Z}\}$ be a wavelet filter. For appropriately chosen N, denote by H_k a matrix of size $(2^{J-k} \times 2^{J-k+1}),\ k = 1, \ldots$ with (i,j)th element

$$h_s, \text{ for } s = (N-1) + (i-1) - 2(j-1) \textit{ modulo } 2^{J-k+1}. \tag{12}$$

Define a matrix G_k as in (12) by using the quadrature mirror filter $\mathbf{g}$. The constant N is a shift parameter and affects the position of the wavelet on the time scale. For the time series $\mathbf{y}$, the following matrix equation (J-step discrete wavelet transformation) gives the connection between $\mathbf{y}$ and the wavelet coefficients $\mathbf{d}$ as in (4)

$$\mathbf{d} = W_J \cdot \mathbf{y},$$

where W_J is defined iteratively,

$$W_1 = \begin{bmatrix} H_1 \\ G_1 \end{bmatrix}, \quad W_2 = \begin{bmatrix} \begin{bmatrix} H_2 \\ G_2 \end{bmatrix} \cdot H_1 \\ G_1 \end{bmatrix},$$

$$W_3 = \begin{bmatrix} \begin{bmatrix} \begin{bmatrix} H_3 \\ G_3 \end{bmatrix} \cdot H_2 \\ G_2 \end{bmatrix} \cdot H_1 \\ G_1 \end{bmatrix}, \ldots$$

The result of a wavelet decomposition of an image is a square with squares-within-squares of low-pass ($\mathcal{H}$) operations and high-pass operations ($\mathcal{G}$) as shown in Figure 8.

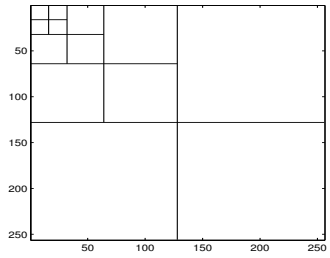


Figure 8: Tessellations for some 2-D wavelet transform of depth 4.

In particular, the first step of the DWT performs the transform on all rows producing two subsets of coefficients as in Figure 9 (left): the left side of the matrix contains downsampled low-pass coefficients of each row and the right contains the high-pass coefficients. The application of the DWT on the columns of the matrix in Figure 9 (left) decomposes an image into a lower resolution approximation image ($\mathcal{HH}_1$) as well as horizontal ($\mathcal{GH}_1$), vertical ($\mathcal{HG}_1$) and diagonal ($\mathcal{GG}_1$) detail components as in Figure 9 (center). To compute a two-level decomposition, the DWT algorithm is again applied on the $\mathcal{HH}_1$ which further decompose the $\mathcal{HH}_1$ part in four subbands $\mathcal{HH}_2$, $\mathcal{GH}_2$, $\mathcal{HG}_2$ and $\mathcal{GG}_2$ (Fig. 9, right).

In general, at each level j, the DWT produces four types of coefficients:

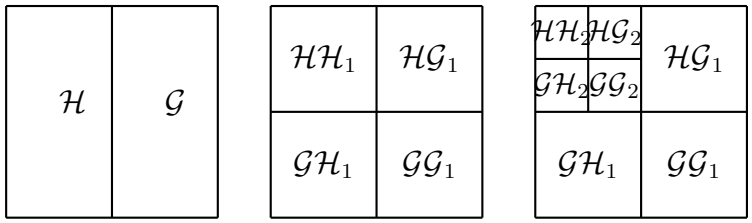


Figure 9: Steps of a two-level discrete wavelet transform of an image: wavelet transform on all rows (left); one-level decomposition (center); two-level decomposition (right).

(a) the coefficients that result from a high-pass filtering g on the rows and columns ($\mathcal{GG}_j$) represent the diagonal features of the image;

(b) the coefficients that result from a a low-pass filtering h on the columns after a convolution with g on the rows ($\mathcal{GH}_j$) correspond to horizontal structures;

(c) the coefficients from high-pass filtering on the rows, followed by low-pass filtering of the columns ($\mathcal{HG}_j$) reflect vertical information, and

(d) coefficients from low-pass filtering h in both directions represent the smoothed image which is further processed in the next step.

The following example illustrates a multivariate wavelet transformation. Consider a gray-scale image A, whose entries a_{ij} correspond to the intensities of gray in the pixel at position (i, j). The process of wavelet decomposition begins by applying the wavelet low pass filter $\mathcal{H}$ and high pass filter $\mathcal{G}$ to the rows of the matrix A. This step produces two matrices $\mathcal{H}_r A$ and $\mathcal{G}_r A$, both of dimension $2^n \times 2^{n-1}$ (the subscripts r suggest that the filters are applied on rows of the matrix A). Next, apply the filters $\mathcal{H}$ and $\mathcal{G}$ to the columns of matrices $\mathcal{H}_r A$ and $G_r A$ obtained from step one, producing matrices $\mathcal{H}_c\mathcal{H}_r A, \mathcal{G}_c\mathcal{H}_r A, \mathcal{H}_c\mathcal{G}_r A$ and $\mathcal{G}_c\mathcal{G}_r A$ of dimension $2^{n-1} \times 2^{n-1}$. The matrix $\mathcal{H}_c\mathcal{H}_r A$ is an average or smooth representation of the original image, while the matrices $\mathcal{G}_c\mathcal{H}_r A, \mathcal{H}_c\mathcal{G}_r A$ and $\mathcal{G}_c\mathcal{G}_r A$ contain detailed features of image A. To produce images with increased smoothness, one may repeat the process using the *average* matrix $\mathcal{H}_c\mathcal{H}_r A$ in place of A. Figure 10 depicts a wavelet-decomposition of the Lenna image, which is used as a standard template in the image processing field, into a smooth part and three detail matrices.

Figure 11 gives another example of a 2-D wavelet transformed image. Note that the wavelet coefficients in the hierarchy below the diagonal emphasize horizontal features in the original image, while the wavelet coefficients in the hierarchy above the diagonal emphasize vertical features in the original image.

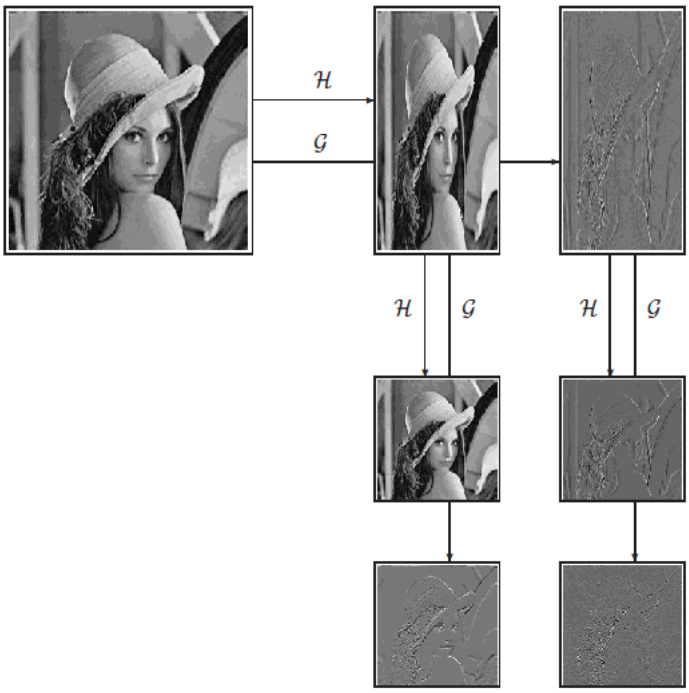


Figure 10: One step in the wavelet transformation of an image. The standard image processing template image Lenna is used.

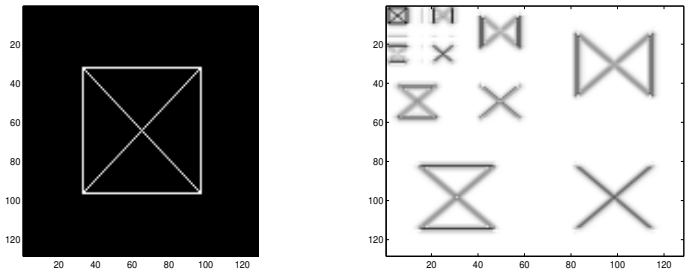


Figure 11: Image of a box with a cross within (left), and its 2D wavelet transform using COIFLET2 basis and $L = 3$ levels of resolutions (right).

1.4 The Scale-Mixing 2-D Wavelet Transform

In this section, we review the notion of the 2-D spectra introduced by Ramirez et al 2011. This relatively new spectra is utilized extensively in the methods of Chapters 2 and 3. Many versions of the 2-D wavelet transform lead to tessellations (or tilling) of the squared image. For example, if instead of (9) and (10), the wavelet atoms are defined as

$$\phi_{(j_1,j_2),\mathbf{k}}(\mathbf{t}) = 2^{(j_1+j_2)/2}\ \phi(2^{j_1}t_1 - k_1, 2^{j_2}t_2 - k_2) \tag{13}$$

$$\psi^i_{(j_1,j_2),\mathbf{k}}(\mathbf{t}) = 2^{(j_1+j_2)/2}\ \psi^i(2^{j_1}t_1 - k_1, 2^{j_2}t_2 - k_2), \tag{14}$$

where i is one of h, v, or d, as in (8) and $(j_1, j_2) \in \mathbb{Z}^2$, then for $X \in \mathcal{L}_2(\mathbb{R}^2)$

$$\begin{aligned} X(\mathbf{t}) &= \sum_{\mathbf{k}} c_{(J_0,J_0),\mathbf{k}}\ \phi_{(J_0,J_0),\mathbf{k}}(\mathbf{t}) \\ &+ \sum_{j>J_0}\sum_{\mathbf{k}} d_{(J_0,j),\mathbf{k}}\ \psi^h_{(J_0,j),\mathbf{k}}(\mathbf{t}) \\ &+ \sum_{j>J_0}\sum_{\mathbf{k}} d_{(j,J_0),\mathbf{k}}\ \psi^v_{(j,J_0),\mathbf{k}}(\mathbf{t}) \\ &+ \sum_{j_1,j_2>J_0}\sum_{\mathbf{k}} d_{(j_1,j_2),\mathbf{k}}\ \psi^d_{(j_1,j_2),\mathbf{k}}(\mathbf{t}), \end{aligned}$$

and a new 2-D wavelet transform, referred to throughout this thesis as the *scale-mixing wavelet transform* is obtained. Notice that (j_1, j_2) in (13) and (14) can be indexed as well as $(j_1, j_1 + s)$, where $s \in \mathbb{Z}$. The new scale-mixing detail coefficients

are defined as

$$\begin{aligned} d_{(J_0,j),\mathbf{k}} &= 2^{(J_0+j)/2} \int X(\mathbf{t})\ \psi^h(2^{J_0}t_1 - k_1, 2^j t_2 - k_2)\ dt_1\ dt_2, \\ d_{(j,J_0),\mathbf{k}} &= 2^{(j+J_0)/2} \int X(\mathbf{t})\ \psi^v(2^j t_1 - k_1, 2^{J_0} t_2 - k_2)\ dt_1\ dt_2, \\ d_{(j_1,j_2),\mathbf{k}} &= 2^{(j_1+j_2)/2} \int X(\mathbf{t})\ \psi^d(2^{j_1} t_1 - k_1, 2^{j_2} t_2 - k_2)\ dt_1\ dt_2. \end{aligned} \tag{15}$$

Similar to the traditional one- and two-dimensional cases, the scale-mixing detail coefficients are linked to the original image (2-D time series) through a matrix equation. Suppose that an $2^n \times 2^n$ image (matrix) A is to be transformed into the wavelet domain. If the rows of A are transformed by a one-dimensional transform given by the $2^n \times 2^n$ wavelet matrix W, then the object WA' represents a matrix in which the columns are transformed rows of A. If the same is repeated on the rows of WA' the result is

$$B = W(WA')' = WAW'. \tag{16}$$

Matrix B will be called the scale-mixing or covariance wavelet transform of matrix A, and will be the basis for defining the scale-mixing spectra. It represents a finite-dimensional implementation of (15) for signal $X(\mathbf{t})$ sampled in a form of matrix A. The term "covariance transform" is motivated by the following fact. If X is a zero mean random vector with covariance matrix A, then $Y = WX$ has covariance $B = WAW'$. Of course, in wavelet transforms, A is arbitrary and not necessarily a covariance matrix (positive definite).

The scale-mixing 2-D transform is operationally appealing. Images analyzed are usually of moderate size, and constructing the appropriate W is computationally fast. Since W is orthogonal, the inverse transform is straightforward,

$$A = W'BW.$$

Unlike the traditional 2-D wavelet transform in which extension to rectangular matrices substantially complicates the algorithm, the corresponding scale-mixing 2-D

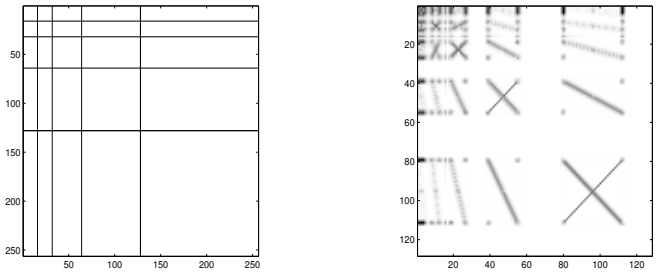


Figure 12: Tessellations for some 2-D scale-mixing wavelet transform of depth 4 (left), and 2-D scale-mixing wavelet transform of the box with cross image (right).

wavelet transforms are straightforward. Since the wavelet transform is applied on the rows first and then on the columns, one can handle not only rectangular images, but also different bases in W and W', multiple transforms $W_1 W_2 A W_2' W_1'$, and so on.

Figure 12 illustrates the 2-D scale-mixing wavelet transform, once again using the box with cross image. By inspecting the tessellation in Figure 12 (left), several hierarchies of detail spaces can be identified. The diagonal hierarchy interfaces coefficients with the same component scales and coincides with the diagonal hierarchy in the traditional 2-D spectra. Just above and below the diagonal hierarchy are hierarchies of detail spaces that interface the scales that differ by 1. For example, the hierarchy above the diagonal, the scales along x-direction are interfaced by the next coarser scale along y-direction. For the hierarchy below the diagonal, roles of x and y are interchanged. Figure 13 (a) shows three hierarchies of detail coefficients: the diagonal hierarchy (circles) and the hierarchies in which dyadic scales differ by 1 (triangles and squares).

Scale-mixing 2-D wavelet transform is typically more compressive than the traditional 2-D wavelet transform, which is a desired property when dimension reduction applications (denoising, compression) are of interest. Informally, if the transform is of depth 2, 9/16 of the coefficients generated by the scale-mixing transform correspond to the differencing filters in two dimensions while for the traditional transform this proportion is only 5/16. The rest of the coefficients correspond to the atoms

containing at least one scaling function. Empirically, the Lorenz curve for squared wavelet coefficients in a scale-mixing transform is typically below the Lorenz curve for the traditional transform of the same depth. For the purpose of spectral analysis, orthogonality is more important than compressibility. The balance of the total energy $E = \text{trace}(AA')$ in the image A, over the scales and mixture of scales is preserved since the orthogonality of W implies,

$$E = \text{trace}(AA') = \text{trace}(BB'),$$

for $B = WAW'$.

1.4.1 Definition of Scale-Mixing Wavelet Spectra

The scale-mixing spectra is defined in terms of the scale-mixing coefficients (15) as

$$S(j) = \log_2 \mathbb{E}\left(d^2_{(j,j+s),\mathbf{k}}\right), \tag{17}$$

where $s \in \mathbb{Z}$ is fixed. The empirical counterpart of (17) is

$$\hat{S}(j) = \log_2\left(\overline{d^2_{(j,j+s),\mathbf{k}}}\right). \tag{18}$$

In (18), $\overline{d^2_{(j,j+s),\mathbf{k}}}$ denotes the average of squared detail coefficients (15) at level $(j, j+s)$. Notice that the case $s = 0$ in (18) corresponds to the traditional diagonal 2-D spectra, see Figure 13.

To calibrate the scale-mixing spectra, consider now a 2-D fBm $B_H(\mathbf{u})$. For such a process, the scale-mixing detail coefficients are given by

$$d_{(j,j+s);\mathbf{k}} = 2^{j+\frac{s}{2}} \int B_H(\mathbf{u})\psi\left(2^j u_1 - k_1, 2^{j+s}u_2 - k_2\right)\, d\mathbf{u},$$

where $\psi \equiv \psi^d$. These coefficients are random variables with zero mean and variance

$$\begin{aligned}\mathbb{E}\left[d^2_{(j,j+s);\mathbf{k}}\right] = 2^{2j+s} \int & \psi\left(2^j u_1 - k_1, 2^{j+s}u_2 - k_2\right) \\ & \times \psi\left(2^j v_1 - k_1, 2^{j+s}v_2 - k_2\right) \mathbb{E}\left[B_H(\mathbf{u})B_H(\mathbf{v})\right] d\mathbf{u}\, d\mathbf{v},\end{aligned} \tag{19}$$

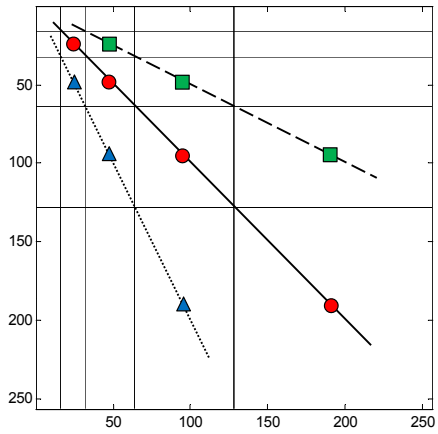


Figure 13: Three detail-space hierarchies generating the scale-mixing 2-D transform, where (j_1, j_2) is indexed as $(j, j+s)$, $s \in \mathbb{Z}$. Circles correspond to $s = 0$, triangles to $s = 1$, and squares to $s = -1$;

(Heneghan et al., 1996; Nicolis et al., 2010). As in Veitch and Abry (1999) and Nicolis et al. (2010) it is assumed here that the coefficients within and across the scales are uncorrelated. This assumption is reasonable; Flandrin (1992) showed that when the number of vanishing moments of the scaling function ϕ is large, the correlation between the coefficients within a scale decays exponentially fast, while the coefficients from different scales are almost uncorrelated.

From (19), it can be seen that

$$\mathbb{E}\left[d^2_{(j,j+s);\mathbf{k}}\right] = 2^{-j(2H+2)}\ V_{\psi,s}(H), \tag{20}$$

where $V_{\psi,s}(H)$ is a constant depending on ψ, H and s, but not on the scale j. A proof of (20) is provided in the Appendix. By taking logarithms in (20),

$$\log_2 \mathbb{E}\left[d^2_{(j,j+s);\mathbf{k}}\right] = -(2H+2)j + \log_2 V_{\psi,s}(H), \tag{21}$$

and thus the Hurst exponent can be estimated from the slope of the linear equation (21). Finally, the empirical counterpart of (21) is a regression defined on

$$\left(j, \log_2 \overline{d^2_{(j,j+s),\mathbf{k}}}\right), \quad s \in \mathbb{Z}. \tag{22}$$

Instead of the sample mean in (22), a more robust location measure could be used, such as the median.

CHAPTER II

CHARACTERIZING EXONS AND INTRONS BY REGULARITY OF NUCLEOTIDE STRINGS

2.1 Introduction

2.1.1 Exons and Introns in Eukaryotic DNA

In all eukaryotic species, a DNA molecule consists of a long double helix of purine nucleotides (denoted as A and G) and pyrimidine nucleotides (denoted as C and T). Genomes of eukaryotic species are generally larger than those of prokaryotes. While the DNA of prokaryotes is gene-rich with a few noncoding regions, eukaryotic DNA contains longer spaces between genes, and genes are partitioned into a great number of regions that will or will not be translated into proteins. Protein synthesis of eukaryotes requires two steps: transcription and translation. During transcription, a pre-mRNA is synthesized from a DNA template. Later, the noncoding regions (introns) are spliced out and the coding regions (exons) are joined to produce mature mRNAs. In the translation step, the mature mRNAs are translated into proteins.

Originally thought to carry unimportant sequences, introns are now known to have biological functions. They harbor a variety of regulatory elements such as untranslated RNAs and splicing control elements, which regulate the splicing process and allow alternative splicing – a mechanism leading to greater variability of gene products (proteins). Specific intron sequences at the exon-intron boundaries and within the intron itself facilitate the pre-mRNA splicing process, as well as the alternative splicing process. In addition to sequence motifs of introns that are driven by their roles during transcription, introns also carry features that associate with exon structure. Zhu and coworkers (2009) analyzed the variability of intron-exon architecture

across many genomes and detected that some intron properties such as length, ordinal position, and GC content are correlated with the exon structure. One notable correlation was observed between the GC content of an intron and its flanking exons (Zhu et al., 2009).

Several studies on the regularity of exons and introns demonstrate that the two types of regions exhibit different degrees of regularity and that regularity patterns can be used to infer the origin or functionality of the sequences. For example, a study of gene origin (Ieviņa et al., 2006), suggests primitive genes – both exon and intron sequences – were highly periodic. However, a sequence's structure changes throughout the course of evolution. Unlike exons, which have coding ability, many intronic sequences lack functional constraints and therefore are free to accumulate a large number of mutations. As a result, introns with no obvious functional roles subsequently lost the sequence regularity pattern (Elder, 2000).

2.1.2 Previous Work on Translating DNA Nucleotides to Numbers

Translating DNA into a numeric form has been done in many ways and has allowed researchers to investigate the properties of protein-coding sequences and noncoding sequences Peng et al. (1992) first mapped nucleotide sequences onto a "DNA walk" in which the walker moves along the DNA sequence, stepping up ($u(i) = +1$) if a pyrimidine occurs and stepping down ($u(i) = -1$) if a purine occurs. They characterize the fractal landscape of DNA quantitatively using the mean fluctuation function $F(l)$, defined by

$$F^2(l) = \overline{[\Delta y(l) - \overline{\Delta y(l)}]^2}, \tag{23}$$

where $y(l)$ for a given l, defined to be $y(l) = \sum_{i=1}^{l} u(i)$, is a trajectory of the DNA walk and $\Delta y(l) = y(l_0 + l) - y(l_0)$, where l_0 are all positions in the gene. In (1), the overline stands for the average. Segments of DNA which are uncorrelated or only short-range correlated have $F(l) \sim l^{1/2}$, while segments with long-range correlation

have $F(l) \sim l^{\alpha}$ ($\alpha \neq 1/2$). For the sequences studied, Peng et al. found long-range correlations (regularity) in intron-containing genes and in non-transcribed regulatory DNA sequences, but not in cDNA sequences or intron-less genes. A similar large scale study has also shown the presence of long-range correlations for noncoding sequences (Buldyrev et al., 1995). More recently, a group of researchers attempted to quantify the degree of non-stationarity of DNA sequences through rescaled range analysis (Boekhorst et al., 2008). They used the rescaled range of a segment to estimate its Hurst exponent, a measure of self-similarity. Their methodology illustrated, in agreement with earlier results, that exons (coding regions) have lower Hurst exponents than introns (noncoding regions).

Other research has suggested more general uses for discovering regularity properties of DNA sequences. Local irregularities along a DNA strand, compared to surrounding regions, have been associated with biological functionality (i.e. coding for proteins and functional RNAs). Haimovich et al. (2006) suggested that if pattern irregularities are observed in introns, it may indicate biological significance of specific intron regions. In addition to the presence/absence of biological functionality, variation of long-range correlation levels of different genomic regions has been used in the study of origin of genes and introns (Ieviņa et al., 2006), and the effects of mutation accumulations or various evolutionary genomic events (replication slippage, recombination, translocation, and transposition) on sequence regularity (Paxia et al., 2002).

Numeric conversion methodologies other than the DNA walk have been proposed. Stoffer et al. (2000) approached the problem of scaling in nucleotide sequences by using so-called "spectral envelopes." The idea behind this methodology is to find numerical scaling values to assign to each category of nucleotide which will maximize the variance of the resulting stationary time series' spectral density across frequencies relative to the total variance. Another more simplistic approach is to represent DNA

with four separate binary indicator sequences corresponding to the four nucleotide bases (Voss, 1992; Afreixo et al., 2004; Yin and Yau, 2007). Often binary indicator sequences are grouped according to their chemical structures for statistical analysis (Cattani et al., 2006; Bai et al., 2007; Arneodo et al., 2011).

One interesting numeric mapping solution uses the concept of symbolic autocorrelation (Pinho et al., 2006). Given the sequence of nucleotide symbols x_i, its symbolic autocorrelation is the numeric sequence r_k

$$r_k = \sum_{i=0}^{n-1} d(x_i, x_{i+k}),$$

where for any two symbols a and b,

$$d(a,b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b \end{cases}$$

Then the discrete Fourier transform of this autocorrelation is the spectrum of the symbolic data.

Our approach in this paper is to translate sequences of DNA into numeric matrices to be analyzed via wavelet analysis. We define the *cumulative evolutionary slope* of a sequence and show how it can be used to assess the scaling in nucleotide sequences. An advantage of the proposed method is its invariance with respect to assignment of nucleotides to their numerical values. Unlike the DNA walks where the nucleotides are assigned numerical values depending on nucleotide characteristics (purine-pyrimidine, weak-strong hydrogen bonds, keto-amino, etc.) and some other spatial assignments, the proposed scaling measure is invariant to the assignment of nucleotides. Thus, subjectivity in numerical translation of nucleotides is eliminated.

2.2 From ACGT to Numbers

2.2.1 Translating to Matrices via Assignment of Unit Vectors

Suppose that a nucleotide sequence of length N is encoded to the index matrix $4 \times N$ such that A is coded as $e_1 = (1,0,0,0)'$, C as $e_2 = (0,1,0,0)'$, G as $e_3 = (0,0,1,0)'$

and T as $e_4 = (0, 0, 0, 1)'$. Denote this matrix with Y. For example,

$$GATCTCT \ldots \longrightarrow Y = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & \ldots \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & \end{bmatrix}.$$

Assume also that N is a power of 2 for implementational purposes. Define Y^* as the matrix formed by accumulating across the rows of Y. Continuing the above example,

$$Y^* = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & \\ 0 & 0 & 0 & 1 & 1 & 2 & 2 & \ldots \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & \\ 0 & 0 & 1 & 1 & 2 & 2 & 3 & \end{bmatrix}.$$

If W_4 is a 4×4 matrix corresponding to Haar wavelet transform of depth 2, then

$$W_4 = \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 & -1/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 & 0 & 0 \\ 0 & 0 & \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix}.$$

Define $D = W_4 \times Y^*$ to be a matrix in which the columns of Y^* are Haar transformed. In D the accumulated unit vectors are replaced by Haar orthogonal vectors that are columns of W_4. This transforms the sparse Y to a more dense representation, D. For example,

$$GATCT \ldots \longrightarrow Y^* \longrightarrow D = \begin{bmatrix} 1/2 & 1 & 3/2 & 2 & 5/2 & \ldots \\ -1/2 & 0 & -1/2 & 0 & -1/2 & \ldots \\ 0 & \sqrt{2}/2 & \sqrt{2}/2 & 0 & 0 & \ldots \\ \sqrt{2}/2 & \sqrt{2}/2 & \sqrt{2} & \sqrt{2} & 3\sqrt{2}/2 & \ldots \end{bmatrix}.$$

Transform the rows of D using wavelet transform that has the depth ≥ 2 to obtain matrix Z of size $4 \times N$. In matrix notation,

$$Z = W_4 \times Y^* \times W'_N,$$

where W_N is an N by N matrix. Now Z is the 2-D scale-mixing wavelet transform of Y^*, see Ramírez-Cobo et al. (2011).

The wavelet basis generating matrix W_N can be arbitrary, but the Haar is most natural since the rows of D are piecewise constant. Also, when N is large (say, $> 2^{11}$) the transformation by W_N is done by Mallat's algorithm instead of direct matrix multiplication.

A submatrix of Z, $Z_2 = Z(3:4, N/2+1:N)$ corresponds to the finest details of scale-mixing 2-D wavelet transform while $Z_1 = Z(2, N/4+1:N/2)$ is the next coarser detail level. Since one dimension of Z is 4, there are only 2 levels of details Z_1 and Z_2 that form the hierarchy for defining the log-spectral slope (see Figure 14 (top)).

Unlike the traditional wavelet log-spectra that is based on log average energies at several levels, usually ≥ 4 depending on size of data, here we have only two spectral points – generated by Z_1 and Z_2, and the slope is estimated from that pair. Log-spectral slope, or simply *slope* s is defined as

$$s = \log(\overline{Z_2^{*2}}) - \log(\overline{Z_1^{*2}}),$$

where $\overline{A^{*2}}$ is the average entry of Hadamard square $A * A$ for arbitrary matrix A, that is, the mean of the squared entries of A. This slope measures the change in energy between adjacent dyadic levels of the transformed matrix. If equal to 0, then the energies are comparable, and this case corresponds to independent random nucleotides. A negative slope indicates presence of regularity, while a positive slope indicates an "explosion of energy" at finer levels of detail, indicating "zig-zaging" irregularities in the sequence.

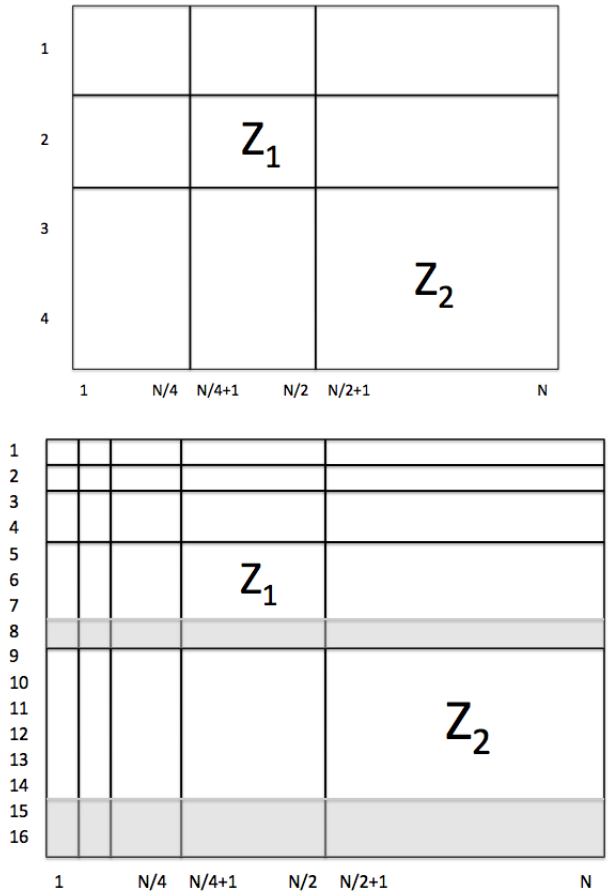


Figure 14: Illustration of submatrices Z_1 and Z_2, the levels of details forming the hierarchy for defining the log-spectral slope, in the original (top) and modified (bottom) procedures.

2.2.2 Equivalence Classes

Consider a particular DNA subsequence, say $ACGT$. Let the sequence be assigned to $\{e1, e2,\ e3, e4\}$ in this order and the resulting slope be s. This sequence consists of two pairs of dinucleotides, AC and GT. The slope is not changed if the nucleotides within each pair are permuted and if the pairs themselves are permuted. For example, the same slope is obtained by assigning $ACTG, CAGT, CATG, GTAC, GTCA, TGAC$, and $TGCA$ to $\{e1, e2, e3, e4\}$.

Furthermore, the 24 permutations of $ACGT$ lead to

$$\frac{4!}{2!\ 2!\ 2!} = 24/8 = 3$$

equivalence classes that result in three different slopes. Table 1 lists the assignments in these equivalence classes.

Table 1: Three equivalence classes of assignments of nucleotides to unit vectors that lead to three different slopes, s_1, s_2 and s_3.

s_1	s_2	s_3
$ACGT$	$AGCT$	$ATCG$
$ACTG$	$AGTC$	$ATGC$
$CAGT$	$GACT$	$TACG$
$CATG$	$GATC$	$TAGC$
$GTAC$	$CTAG$	$CGAT$
$GTCA$	$CTGA$	$CGTA$
$TGAC$	$TCAG$	$GCAT$
$TGCA$	$TCGA$	$GCTA$

We notice that each class represents a nucleotide characteristic. The class of AC and GT corresponds to amino and keto, that of AG and CT purines and pyrimidines, and that of AT and CG weak hydrogen bonds and strong ones.

2.2.3 Invariant Translation Procedure

The slope should not depend on the way in which nucleotides are assigned to unit vectors. The simplest way to find a representative slope is to average the slopes s_1, s_2 and s_3, where the subscript denotes the equivalence class. An alternative (and better) way is to assign nucleotides to unit vectors based on representatives from each of the equivalence classes and stack these unit vectors together. For example, $ACGT$, $AGCT$, and $ATCG$ could each be assigned to (e_1, e_2, e_3, e_4), or equivalently, any other representative triple from the three columns in Table 1. Ultimately, each nucleotide is assigned a vector of length 12. For example, if $ACGT$, $AGCT$, and $ATCG$ are used, C would correspond to $(0\ 1\ 0\ 0\ \ 0\ 0\ 1\ 0\ \ 0\ 0\ 1\ 0)'$.

Following this procedure, Y becomes a $12 \times N$ matrix. In implementations, a matrix Y with 16 rows is used where the last four rows are arbitrary (say zeros) and serve only to fulfill the power of 2 requirement for operational use of wavelets. Since

the Haar basis is used, this padding does not leak to the relevant coordinates. Then Y^* is computed by accumulating across the columns of Y. If Z is produced using the Haar wavelet, and the slope is estimated, this slope is invariant with respect to permutation of coding. In matrix notation,

$$Z = W_{16} \times Y^* \times W'_N.$$

In this case, the submatrices of Z defining the slope are $Z_2 = Z(9:14, N/2+1:N)$ and $Z_1 = Z(5:7, N/4+1:N/2)$ (see Figure 14 (bottom)). This resulting slope is invariant with respect to the assignment of unit vectors e_1 through e_4 to the nucleotides, as long as the assignments from each of the three equivalence classes are used to form matrix Y.

2.2.4 Cumulative Evolutionary Slope

Take a submatrix Y^*_k of size $16 \times k$ and shift it along the nucleotide. Transform it to Z_k as

$$Z_k = W_{16} \times Y^*_k \times W'_k,$$

and find corresponding slopes. This sequence of slopes is the *cumulative evolutionary slope* for the sequence. If the Haar wavelet basis is used, the shifts should be with steps divisible by 4. Values of k that are too small lead to noisy evolutionary slope, while values too large lead to loss of locality. As an illustration of the calculation of the cumulative evolutionary slope, Figure 15 shows three shifts in a nucleotide sequence generate matrix Z_k and associated slopes.

2.3 Application

To illustrate our methodology, we study regularity characteristics of exons and introns from the honeybee's first chromosome. The average GC contents for the sequences analyzed are 32.16% for exons and 24.59% for introns, and the average exon and

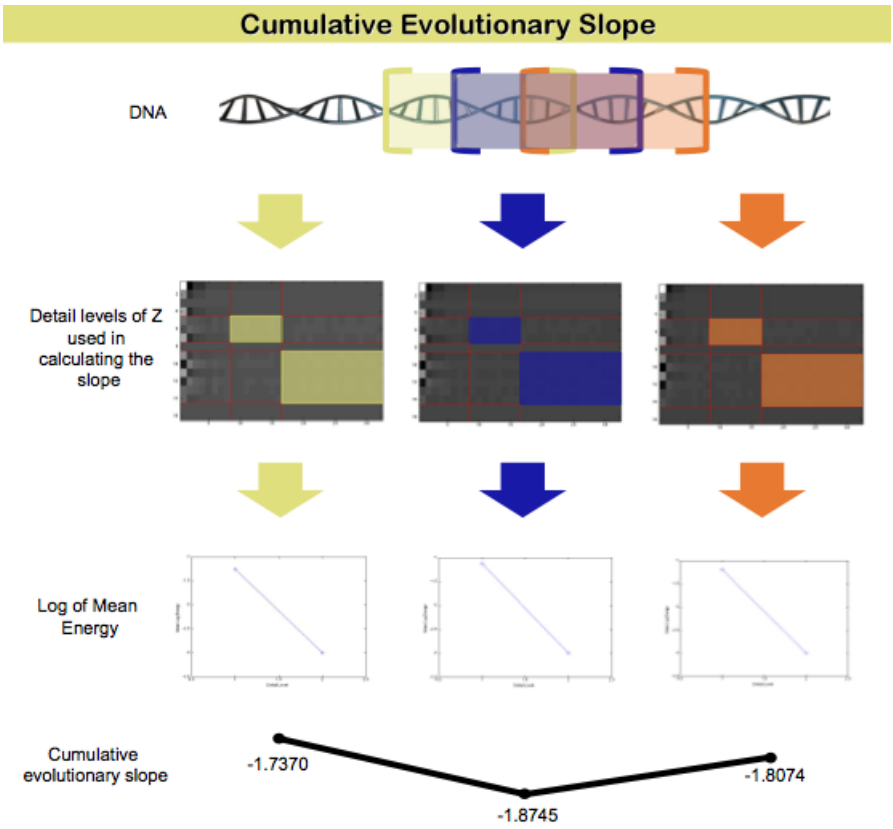


Figure 15: Cumulative evolutionary slope calculation. Overlapping sequences of DNA nucleotides are represented as matrices, the scale-mixing wavelet transformation is applied to these matrices, log average energies are computed for the shown detail levels, and slopes are calculated.

intron lengths are 239 nt and 1,791 nt respectively. A comprehensive study of the honeybee genome indicates honeybee genes are much depleted in C and G nucleotides (gene-averaged GC content of 29%) (The Honeybee Genome Sequencing Consortium, 2006).

2.3.1 Data

The reference DNA sequence from chromosome 1 (linkage group LG1) of the honeybee representative strain, *Apis mellifera* Amel_4.5, was used in our analysis. We obtained the sequence from the NCBI Genome Database (`http://www.ncbi.nlm.nih.gov/genome`). The DNA sequence of chromosome 1 (NC_007070.3) contains a total of 29,893,408 base pairs. The overall GC content

is 31.20%.

For each gene, we used the NCBI annotations of coding sequences (CDS) to identify exon-intron boundary. Then, we parsed the gene sequences into coding and noncoding regions. There are 1,669 genes in the honeybee first chromosome, but only 376 genes contain known locations of CDS. Therefore, we limited our analysis to this subset of genes. DNA has a double helix structure with a complementary nucleotide sequence on each strand. To represent the duplex structure as a single strand, we treated all gene sequences on the reverse strand as if they were located on the forward strand. Specifically, we corrected the gene direction and presented the gene sequences with their complementary sequences. This way, every gene can be read from left to right during the DNA analysis.

As is true with many eukaryotic genomes (Deutsch and Long, 1999; Sakharkar and Chow, 2004; Zhu et al., 2009), honeybee introns are much longer than their exons. A recent study on GC content indicates two possible evolutionary mechanisms for discriminating exons from introns (Amit et al., 2012). Through the first mechanism, low GC content exons and lower GC content shot introns evolve into a DNA strand of comparable GC content throughout while maintaining the short intron length. Through the second mechanism, exon-intron GC content is preserved but intron length is increased. The overall low GC content and the long introns of the first chromosomal honeybee DNA may imply the low GC-long introns hypothesis (Amit et al., 2012).

2.3.2 Comparing Regularity of Honeybee and Simulated DNA

To assess regularity characteristics of the honeybee DNA, we first compute global slope measures. Take as an example a single gene sequence from the honeybee's first chromosome:

```
dna = [TCGTGAAGAGGCAAAGGAATCAATAAACGAAGTTGCGGTGAATAGCGA...
```

...ATCCACTGGGCCGGATATTTATCACGTCCCTCGTGTCCACTTTCAAAG]

As the length of this sequence is 877 nucleotides but computing global slope requires the length to be a power of 2, we truncate the sequence to include the first 512 nucleotides and calculate the global slope of this shortened sequence. We also generate 10,000 random DNA-like sequences, maintaining the proportions of nucleotides from the original gene sequence (in this case: 33.03% A, 16.28% C, 19.84% G, 30.85% T). These simulated DNA-like sequences serve as a control for examining overall regularity characteristics of the actual DNA sequence. By holding the proportions of nucleotides constant, we remove any effects of GC content on the difference in regularity characteristics. Figure 16 plots the bootstrap distribution of global slopes from simulated DNA-like sequences as a histogram and plots the global slope from the honeybee DNA sequence as a vertical red line. Comparing the actual and simulated DNA sequences, the actual DNA sequence's slope falls in the left tail of the control distribution. This result indicates that the honeybee nucleotide sequence is generally more regular than the randomly generated DNA based on the bootstrap distribution of the slopes from DNA-like sequences.

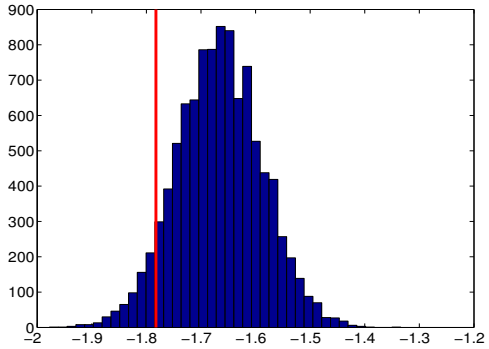


Figure 16: Global slope for the honeybee DNA (red line at -1.7825) sequence and empirical distribution of slopes for 10,000 simulated random DNA-like sequences of length 2^9.

Figure 17 compares the cumulative evolutionary slope of the real DNA to that of a simulated random DNA-like sequence. Note that this time, the gene sequence only

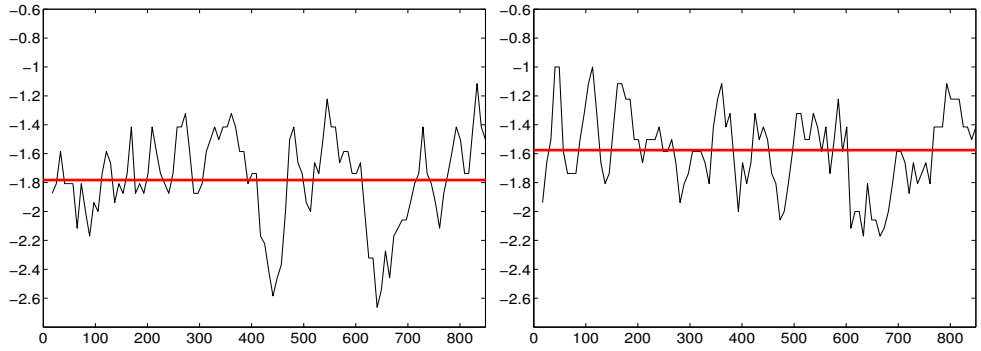


Figure 17: (Left) Honeybee cumulative evolutionary slope for gene "LOC100577807"; (Right) Cumulative evolutionary slope for a random DNA-like sequence; In both cases window size was 2^5

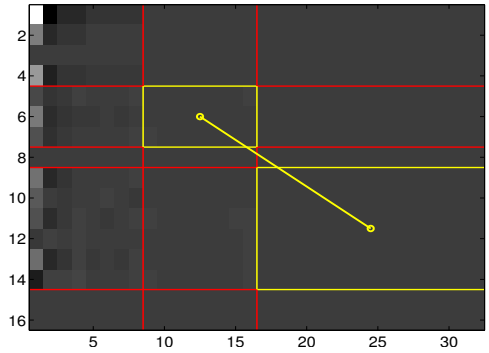


Figure 18: An illustration of detail levels of Z used in the slope calculation for window size 32, honeybee DNA.

has to be truncated so that the length is divisible by the step size. In each plot, the red horizontal line indicates the overall mean of the slopes. The average cumulative evolutionary slope for the honeybee gene is around -1.8, while the average cumulative evolutionary slope for the simulated sequence is around -1.6. These plots support the claim that the actual honeybee DNA is more regular than the randomly generated DNA.

Figure 18 shows Z, the 2-D scale mixing wavelet transform of Y^*, with highlighted detail levels used in the local slope calculation. As previously stated, rows 8, 15, and 16 are excluded in the slope calculation due to the arbitrary rows of zeros added to the original numeric matrix in order to satisfy the power of 2 requirement for operational use of wavelets.

2.3.3 Comparing Regularity of Exons and Introns

What happens when we compare sequences of exons and introns within the honeybee? Considering 376 genes for which the coding designations are known, we plot the cumulative evolutionary slope for each gene sequence. We label sections of the plot as introns, exons, or a combination of the two. We find it necessary to define combination regions due to our evolutionary slope methodology. Since we calculate local slopes for overlapping regions of DNA nucleotides of fixed length, some regions nucleotides contain both exons and introns. Without defining combination regions, the local slopes of exons and introns are averaged together, causing ambiguity in the results.

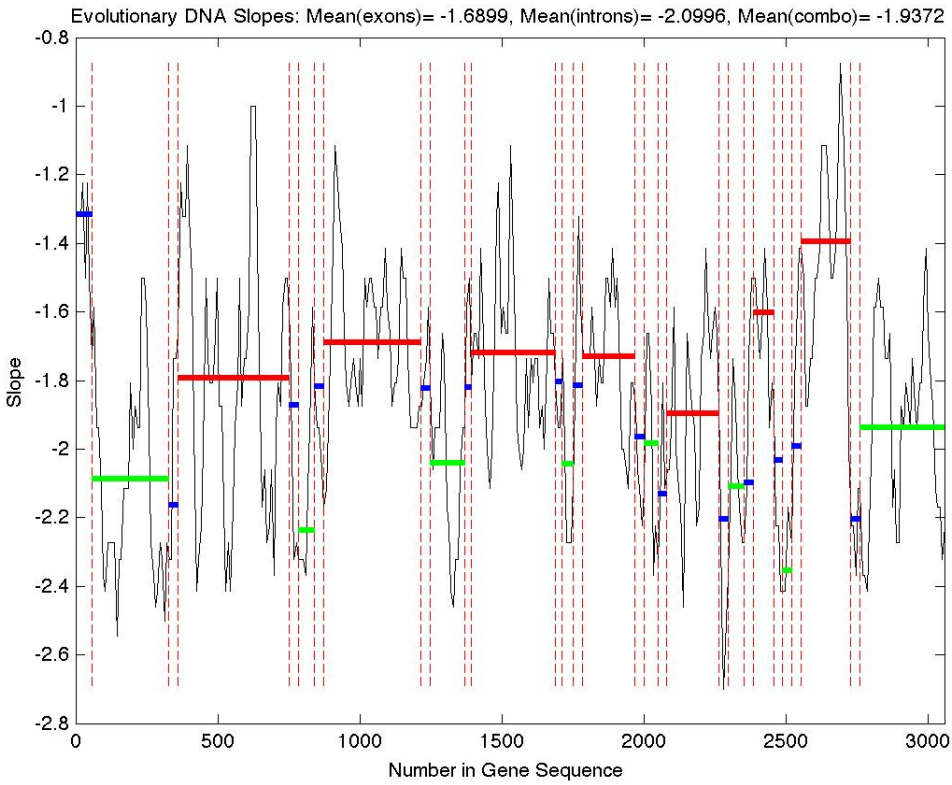


Figure 19: Honeybee cumulative evolutionary slope for gene "LOC408625" on the first chromosome. Solid red line: Average slope for a coding sequence (exons), solid green line: Average slope for a noncoding sequence (introns), solid blue line: Average slope for a sequence with a mixture of coding and noncoding, dotted red line: Division between type (exons, introns, combination) of region.

Figure 19 shows the cumulative evolutionary slope for one of the genes on the first chromosome. The dotted red lines show divisions between types of regions (exons, introns, combination). Solid red lines indicate average slope values for exons, solid green lines indicate average slope values for introns, and solid blue lines indicate average slope values for sequences with a mix of exons and introns. We choose a window size of 2^5 with step size 2^3 to capture characteristics of coding sequences, which are sometimes only around 50 nucleotides in length.

It is evident from Figures 19 and 20 that exons are associated with less negative slopes (more irregular) while introns are associated with more negative slopes (more regular). Figure 19 gives an example of results from a shorter gene on the first chromosome, while Figure 20 gives two examples of results from longer genes with long introns (characteristic of the honeybee genome, as discussed previously) on the first chromosome. Similar results are obtained for the other genes on the first chromosome. The average cumulative evolutionary slope across all 376 genes for exons and introns are -1.7395 and -1.8410, respectively (sample sizes: 1974, 2296). A two sample t-test reveals this difference in slopes to be highly significant (t-statistic=17.4, p-value ≈ 0). Note that some coding, noncoding, or combination regions were excluded due to infinite slopes. These infinite slopes are due to zero-valued mean energies for the finest detail level (see Figure 18). In our analysis, 69 genes have at least one section with infinite slope. Interestingly, of those 69 genes containing at least one infinite slope, 57 have infinite slopes for intron sections only.

To better understand the level of separation between the slopes of exons and introns, we find kernel density estimates computed at 100 points covering the range of the data (Figure 21). In addition, we compute the slope value ($s^* = -1.735$) to discriminate between exons and introns which maximizes the Youden Index, defined as

$$J = sens + spec - 1,$$

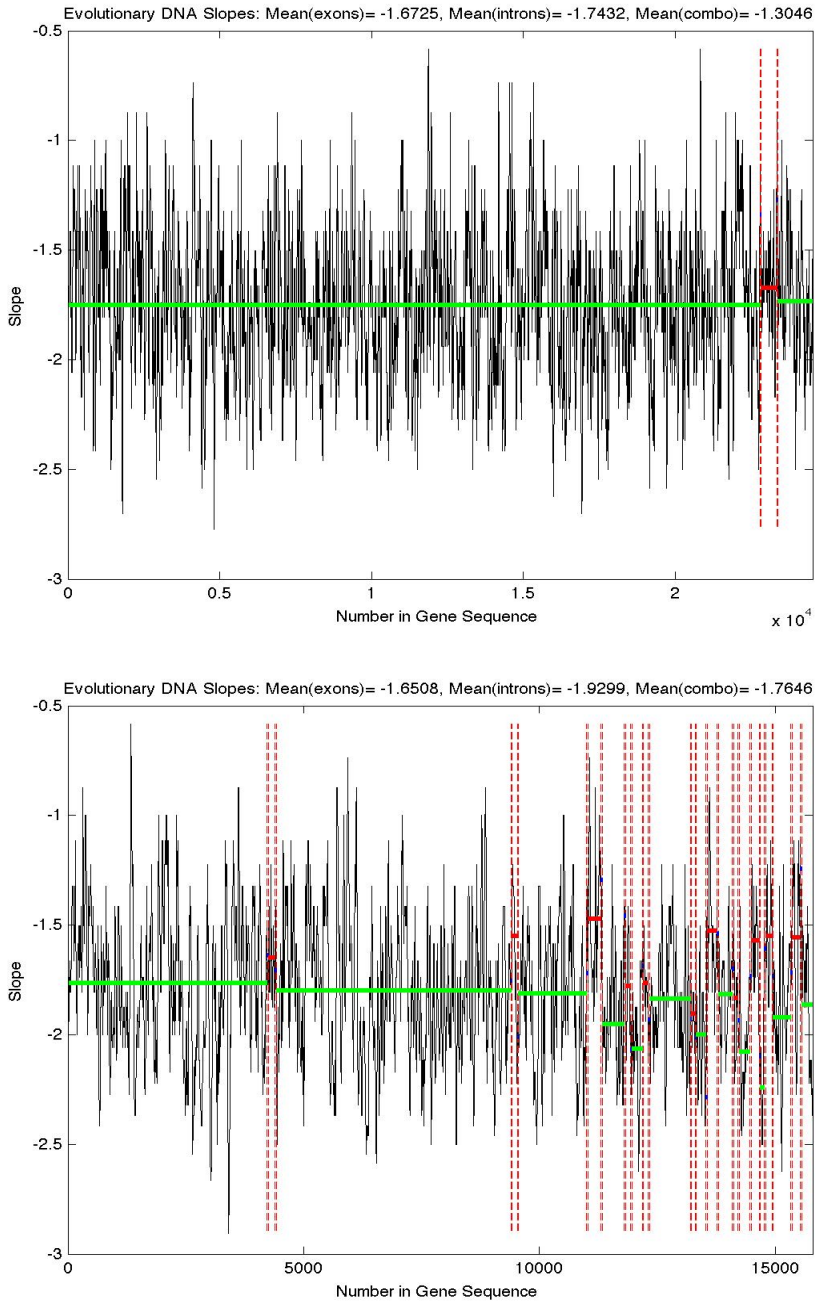


Figure 20: Honeybee cumulative evolutionary slope for genes "Ard1" and "Dat" on the first chromosome. Solid red line: Average slope for a coding sequence (exons), solid green line: Average slope for a noncoding sequence (introns), solid blue line: Average slope for a sequence with a mixture of coding and noncoding, dotted red line: Division between type (exons, introns, combination) of region.

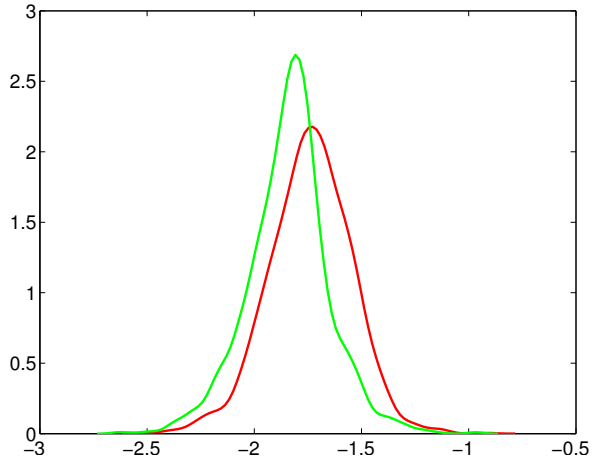


Figure 21: Kernel density estimates for cumulative evolutionary slopes of exons (red) and introns (green)

where *sens* is the sensitivity and *spec* is the specificity of the classification. Table 2 summarizes the classification results for $s^* = -1.735$. The sensitivity and specificity of this classification are 50% and 76.2%, respectively, and the overall accuracy is 64.1%. The Matthews correlation coefficient (MCC) is 0.27. Our intention in presenting these classification results is not to suggest that this procedure is meant for classifying sequences of nucleotides as either exons or introns. Rather, our intention is to present the level of separation between the cumulative evolutionary slopes of exons and introns, as a mechanism for characterizing their regularity properties.

Table 2: Classification results for the cut-off slope value, $s^* = -1.735$, which maximizes the Youden Index.

	Exons	Introns	Total
Slope $> s^*$	987	546	1533
Slope $\leq s^*$	987	1750	2737
Total	1974	2296	4270

2.4 Discussion

We have proposed a new method for representing sequences of DNA nucleotides as numeric matrices in order to analytically investigate regularity characteristics of DNA. Previous methods, where the nucleotides are assigned numerical values depending on nucleotide characteristics (purine-pyrimidine, weak-strong hydrogen bonds, keto-amino, etc.), lead to different regularity measures when different assignments are used. Our proposed method, however, results in a consistent regularity measure through the use of equivalence classes in forming the assignment procedure. Thus, subjectivity in numerical translation of nucleotides is eliminated.

We have also defined the *cumulative evolutionary slope* as a sequence of log-spectral slopes computed from submatrices of wavelet transformed matrices corresponding to overlapping sequences of DNA nucleotides. Shorter overlapping sequences result in noisier cumulative evolutionary slope, while longer overlapping sequences result in smoother cumulative evolutionary slope.

In order to illustrate our methodology, we have analyzed 376 genes from the first chromosome of the honeybee. We found that introns are significantly more regular (lead to more negative slopes) than exons, which agrees with the results from the literature where regularity is measured on "DNA walks." Based on the work of Ieviņa et al. (2006) and Elder (2000), the presence of long-range correlations within honeybee introns suggests their modern introns may be almost identical to the primordial ones and/or the introns were subjected to large evolutionary constraints, thus maintaining their primitive structure.

This hypothesis is supported by well studied organizational and evolutionary characteristics of the honeybee genome. The honeybee and two other insects, the fruitfly and the malaria mosquito, share about one thousand ancient genes (The Honeybee Genome Sequencing Consortium, 2006). This gene set was used to identify the honeybee's evolutionary rate from its vertebrate ancestors. The study shows that the

honeybee retains the greatest fraction of ancient genes ($\approx 33\%$) and ancient introns ($\approx 80\%$). It highlights that the honeybee's genes appear to be ancient and that the honeybee evolves more slowly than the fly and the mosquito (The Honeybee Genome Sequencing Consortium, 2006). Therefore, our findings that the honeybee's genes have low GC content and long introns with high regularity are in concordance with the high proportion of ancient genes and ancient introns.

Long introns are also an indication of regulation complexity, since long introns can accommodate a larger amount of regulatory sequences (Vinogradov, 2006). Long introns are often observed in genes that require sophisticated regulatory mechanisms, such as tissue-specific genes or intermediately expressed genes. Due to the limited number of genome annotations for the honeybee, we do not have complete knowledge of the gene functions. However, we speculate that many of the genes in our dataset are tissue-specific or have differential expression among castes. Therefore, it is possible that the honeybee's introns are embedded with various regulatory elements. Nonetheless, our novel mathematical approach for studying the sequence regularity of the honeybee's first chromosome elaborates interesting patterns on exonic and intronic regions, which can be related to their origin and functions.

CHAPTER III

WAVELET-BASED SCALING INDICES FOR CANCER DIAGNOSTICS

3.1 Ovarian Cancer Diagnostics

3.1.1 Introduction

The National Cancer Institute estimates around 22,000 new cases of ovarian cancer in the United States in 2014. While this number is small in comparison with prevalence statistics for other forms of cancer, ovarian cancer has a particularly low survival rate (estimated 44.6%). Unlike breast cancer or cervical cancer, which can be detected by mammograms and pap tests, there is no formal screening exam for ovarian cancer. Rather, women experiencing certain symptoms are encouraged to see their doctors to receive thorough physical exams, blood tests, and/or ultrasounds to assess the presence of ovarian cancer. However, there are often no symptoms in the early stages of the disease. Even when symptoms do occur, they tend to be nonspecific, delaying the investigation for ovarian cancer. Due to the difficulty of detecting it in its early, treatable states, ovarian cancer has been called the "silent killer."

Ovarian cancer is a disease produced by the rapid growth and division of one of the three major cell types present in the ovary: germ cells, stromal cells, and epithelial cells. Most ovarian cancer research focuses on epithelial ovarian cancer, the most common ovarian malignancy. Previous work has evaluated the use of gene mutations, protein biomarkers, and metabolomic biomarkers in ovarian cancer screening (Berzina et al., 2013; Williams et al., 2007; Guan et al., 2009). Thus far, an assay of the glycoprotein CA125 is the only FDA-approved blood test for the detection of epithelial ovarian cancer. However, it offers little utility in early stage detection. Furthermore,

classification based on CA125 is limited by a lack of sensitivity (50%-60% for early disease detection) (Williams et al., 2007).

Guan et al. (2009) used support vector machines in order to classify cases and controls using serum samples, achieving over 90% accuracy. These results show promise that this kind of approach may lead to the development of an accurate and reliable metabolomic-based approach for detecting ovarian cancer. However, the Guan et al. (2009) study seems to be the only one of its kind to date. Therefore, in this study, we further investigate the use of metabolic data for detecting ovarian cancer.

Traditionally, these types of spectrometry data are analyzed by investigating peaks in the mass spectra, guided by expert knowledge. However, our wavelet-based scaling approach has several advantages. Overall, our approach utilizes each complete set of spectrometry data rather than only data relating to peaks, allowing additional information to be used in classifying cases and controls. In particular, when trends in data are irrelevant and when smoothing does not make sense, scaling analysis of noisy measurements has been shown to yield useful information. For example, in a study on links between changes in pupil diameter and ocular pathologies, Shi et al. (2006) argue that trends in high frequency measurements (> 200 Hz) are irrelevant since they could be affected by the change of environmental light intensity, clearly not related to the pathologies. However, the scaling in these measurements assessed by the Hurst exponent carries discriminatory information about the eye pathologies. Similarly, Jung et al. (2010) propose predicting cysteine concentration of human plasma using scaling of spectroscopy data, since traditional analysis of the H NMR spectra of human plasma can be considered irrelevant to the cysteine concentration because the dominant spectral measurements are insensitive to directly detect cysteine.

Another important property of scaling is that it is invariant with respect to shift/scale of the spectra, and does not require data preprocessing steps such as baseline correction, peak alignment and normalization. The consequence is that Hurst

exponent/slope estimators are robust with respect to changes in a few dominant resonance intensities corresponding to expressed metabolites or marker chemicals.

If a signal has high Hurst exponent, the autocorrelations (correlations between the signal and its shifts) are strong, signifying considerable internal regularity. On the other hand, a signal with low Hurst exponent exhibits intrinsic irregularity and antipersistency. In terms of spectrometry data, spectra with a larger Hurst exponent would possess more internal regularity and autocorrelation. This informally means that metabolites communicate more when the Hurst exponent is higher and that they are more "co-expressed."

3.1.2 Data

The data was obtained through the School of Biology at Georgia Institute of Technology. This sample was run on the LC(-)/MS/MS platform and consists of 40 cases and 40 controls. Stages of the cancer for the cases are III and IV, with most in stage III. Participant age ranges from 50 to 73 years, with an average of around 61 years.

Liquid chromatography mass spectrometry (LC-MS) is an high performance liquid chromatography (HPLC) system with a mass spectrometry detector. First, the HPLC separates chemicals by conventional chromatography on a column. Usually the method will be reverse phase chromatography, where the metabolite binds to the column by hydrophobic interactions in the presence of a hydrophilic solvent (for instance water) and is eluted off by a more hydrophobic solvent (methanol or acetonitrile). As the metabolites appear from the end of the column they enter the mass detector, where the solvent is removed and the metabolites are ionized. The metabolites must be ionized because the detector can only work with ions, not neutral molecules. The mass detector then scans the molecules it sees by mass and produces a full high-resolution spectrum, separating all ions that have different masses. In tandem mass spectrometry (MS/MS), two mass spectrometers are used in series.

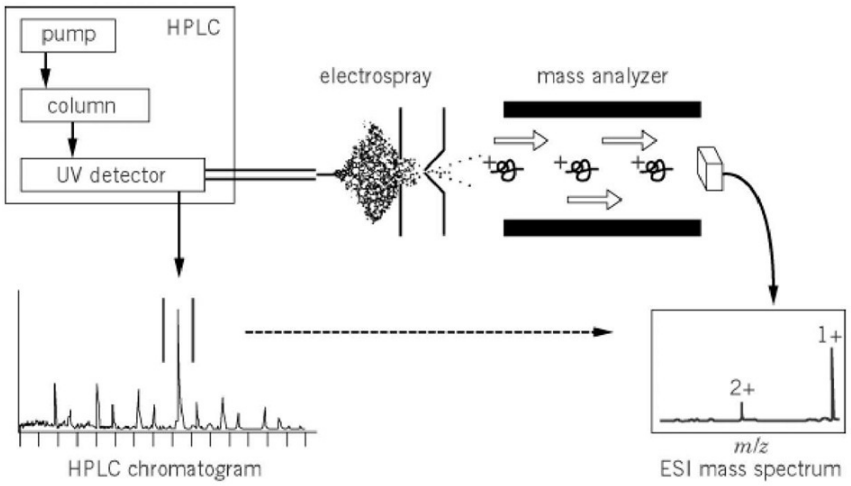
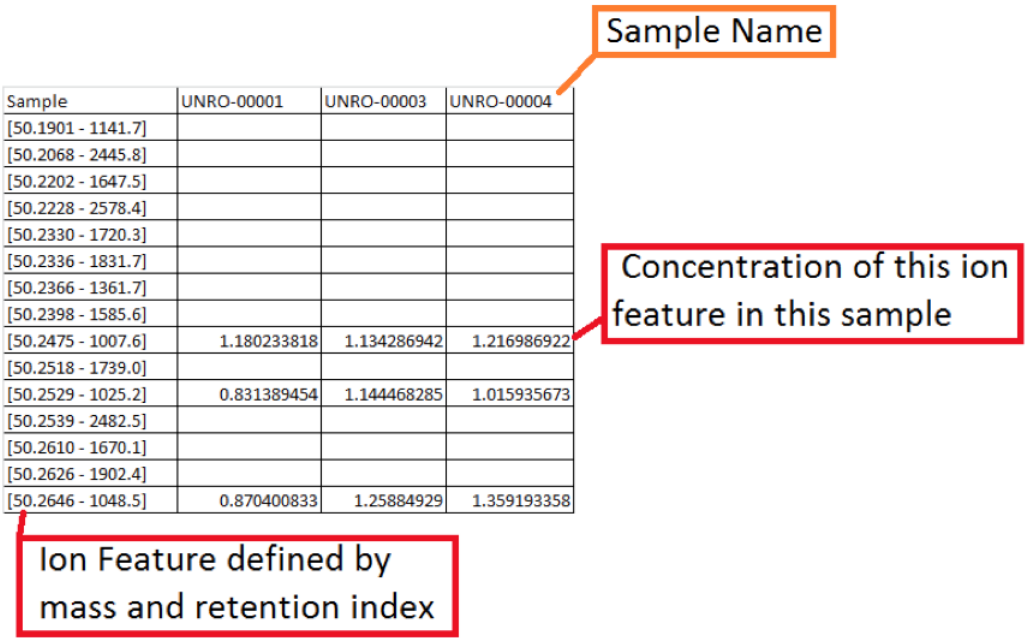


Figure 22: Illustration of LC-MS.

An ion feature is characterized by mass, retention index and the ion chromatographic peak area. To find out the ion features among the samples, ion features in samples are binned by mass and retention index. The average mass and retention index of ion features binned from samples are the mass and retention index of the common ion features among the samples. To account the instrument variability, the concentration of ion features in each sample is normalized, taking the median concentration as 1. Figure 23 shows the format of the data.



Sample	UNRO-00001	UNRO-00003	UNRO-00004
[50.1901 - 1141.7]			
[50.2068 - 2445.8]			
[50.2202 - 1647.5]			
[50.2228 - 2578.4]			
[50.2330 - 1720.3]			
[50.2336 - 1831.7]			
[50.2366 - 1361.7]			
[50.2398 - 1585.6]			
[50.2475 - 1007.6]	1.180233818	1.134286942	1.216986922
[50.2518 - 1739.0]			
[50.2529 - 1025.2]	0.831389454	1.144468285	1.015935673
[50.2539 - 2482.5]			
[50.2610 - 1670.1]			
[50.2626 - 1902.4]			
[50.2646 - 1048.5]	0.870400833	1.25884929	1.359193358

Figure 23: Illustration of the mass spectrometry data.

3.1.3 Time-varying Slope

A signal of length 67,589 represents each of the metabolic characteristics of cases and controls. Examples are shown in Figures 24 and 25. The signals are then centered and accumulated for increased stability.

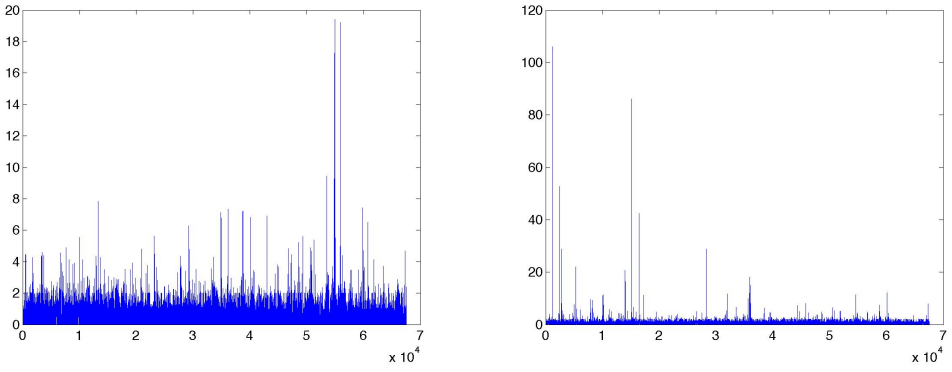


Figure 24: Examples of ion feature signals for cases.

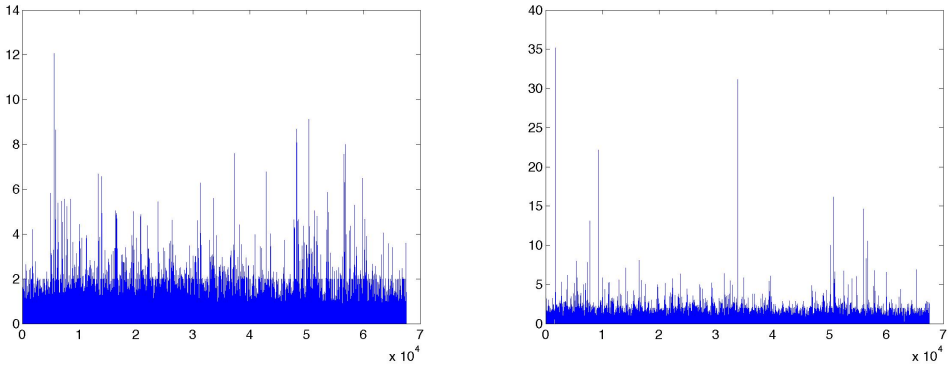


Figure 25: Examples of ion feature signals for controls.

A time-varying slope measure is computed as follows: Take a sub-signal of length 2^{10} (window size) and shift it along the entire signal, with step size 2^5. Then, wavelet transform each sub-signal using the Haar filter ($[\frac{\sqrt{2}}{2}\frac{\sqrt{2}}{2}]$) and weight the resulting wavelet coefficients at each detail level using Gaussian kernel weighting. By using this kernel weighting,

$$\frac{\sum_j w_j d_{ij}^2}{\sum_j w_j}$$

is used in place of $\overline{d_{ij}^2}$ to represent the energies in calculating the log-spectral slope, where d_{ij} represents the wavelet coefficient at the ith level of detail in the jth position. Finally, compute log-spectral slopes for each sub-signal and plot the results. The Gaussian kernel weighting yields more localized slope estimates while providing smoother series of time-varying slopes than would be achieved by simply using a smaller window size.

Figure 26 shows the results for all cases and controls. The time-varying slope series are plotted on top of each other, with cases plotted in blue and controls plotted in green. Plotting the series in this way leads to several observations: (1) The time-varying slope series of cases seem to have higher variability than those of controls. (2) There are specific positions in the signal where the local slope seems particularly discriminatory.

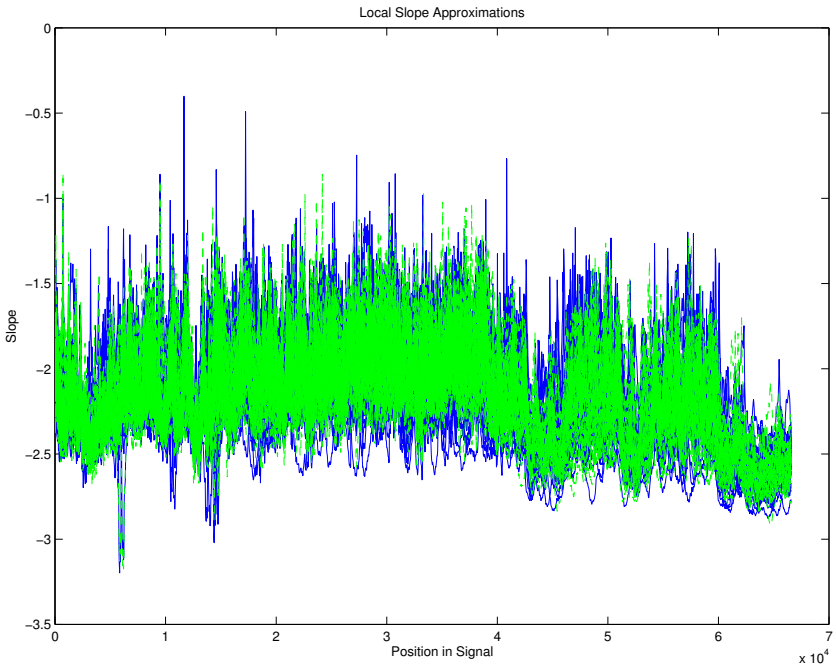


Figure 26: Time-varying log-spectral slopes of ion feature signals for cases (blue) and controls (green).

3.1.3.1 Classification Based on Time-Varying Slopes

Input variables to the classification procedure are based on observations (1) and (2) above. Two variables represent the median time-varying slopes of an ion feature signal within each of the highlighted regions shown in Figure 27. A measure of central tendency is chosen since in both regions, cases seem to generally have less negative slopes than controls. The median is chosen as a more robust alternative to the mean. Another input variable is the coefficient of variation (CV) of the slopes for each signal, or

$$\frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(s_i-\bar{s})^2}}{\bar{s}},$$

where $s_i, i = 1, 2, ..., 2081$ is the slope computed for window i of size 2^{10}.

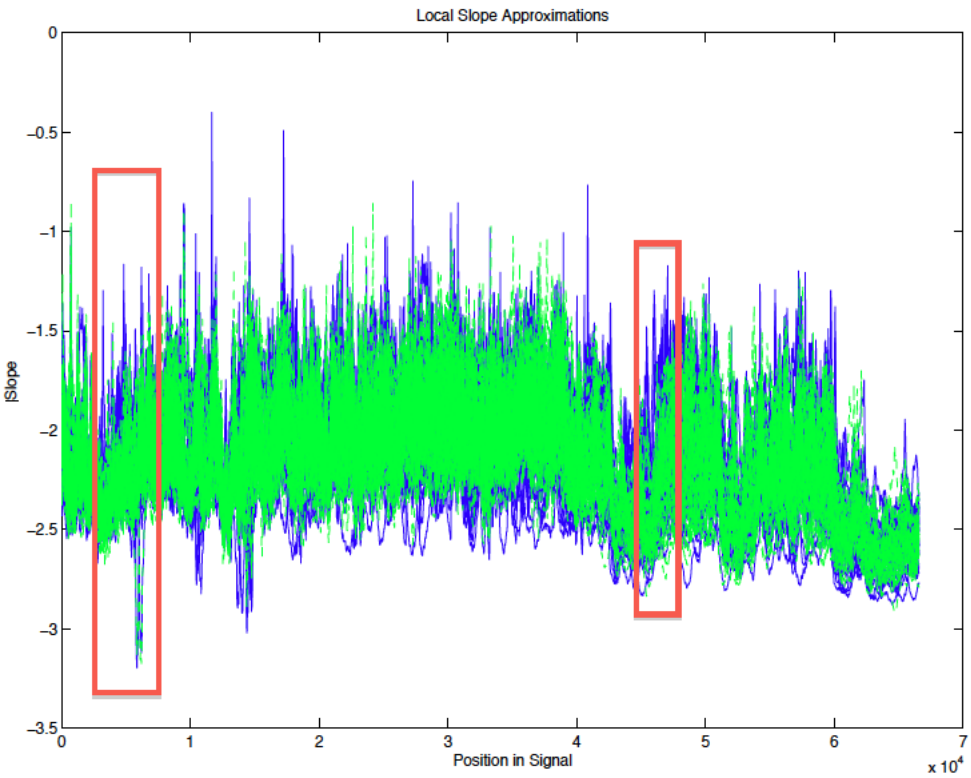


Figure 27: Regions of time-varying log-spectral slopes used to create classification inputs.

The results of 1,000 iterations of the classification procedure using these variables are summarized in terms of sensitivity, specificity, and overall accuracy rate. For each iteration, the data set is split into a 70% training set (60 signals) and a 30% testing set (20 signals). Then, the three time-varying slope related variables are used as features to train a support vector machines (SVM) model with linear kernel. Table 3 summarizes the results, assessed on the testing set (mean accuracy = 64.59%).

3.1.4 Phase

Using a Daubechies complex orthonormal wavelet filter, both real and imaginary wavelet coefficients are obtained for each of the 2^{10} length sub-signals from the time-varying slope analysis. Then, phase information for each window is calculated by

$$\phi = \arctan\left(\frac{d_I}{d_R}\right),$$

where d_I is an imaginary wavelet coefficient and d_R is a real wavelet coefficient. Phase information is calculated for all possible dyadic levels.

3.1.4.1 Classification Based on Phase

Another procedure for classifying ovarian cancer cases and controls is based on the phase information obtained. As before, the data set is split into a 70% training set (60 signals) and a 30% testing set (20 signals). For each signal, at each possible dyadic level, the CV of the phase is computed. Then, various phase CV thresholds are considered for classifying cases and controls. Figure 28 shows the receiver operating characteristic (ROC) curve corresponding to classification procedures at different threshold values. The chosen phase CV threshold of 3.9268 is the one which maximizes the Youden Index,

$$J = \frac{sens + spec - 1}{\sqrt{2}},$$

where *sens* is the sensitivity and *spec* is the specificity of the classification. Table 3 summarizes the results, assessed on the testing set (accuracy = 61.57%).

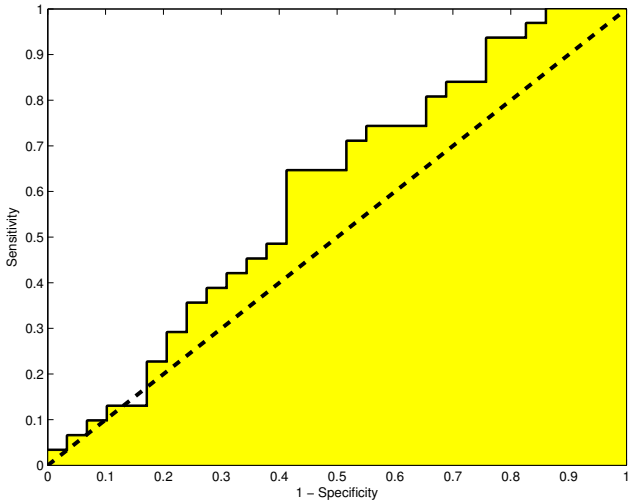


Figure 28: ROC curve: True positive rate against false positive rate as the classification threshold is varied.

3.1.5 Combining Classification Procedures

We try combining the classification procedure based on time-varying slopes and the classification procedure based on phase in two ways: in serial and in parallel. For the combined serial classification, both individual classification procedures must label an observation as a case in order to be labeled overall as a case. Otherwise, the observation is labeled as a control. For the combined parallel classification, an observation labeled as a case by either individual classification is labeled as a case overall. An observation labeled as a control by both individual classification procedures is labeled as a control overall. Figures 29 and 30 demonstrate the difference between these two combination approaches. In general, procedures combined serially lead to higher specificity and lower sensitivity overall, while procedures combined in parallel lead to higher sensitivity and lower specificity overall.

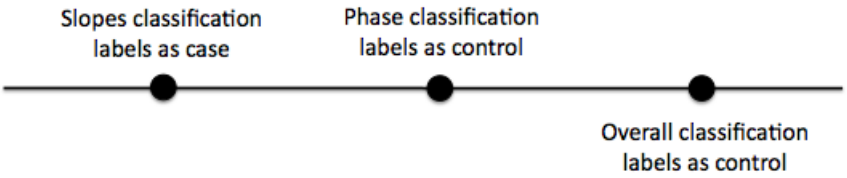


Figure 29: Combining classification procedures: serial.

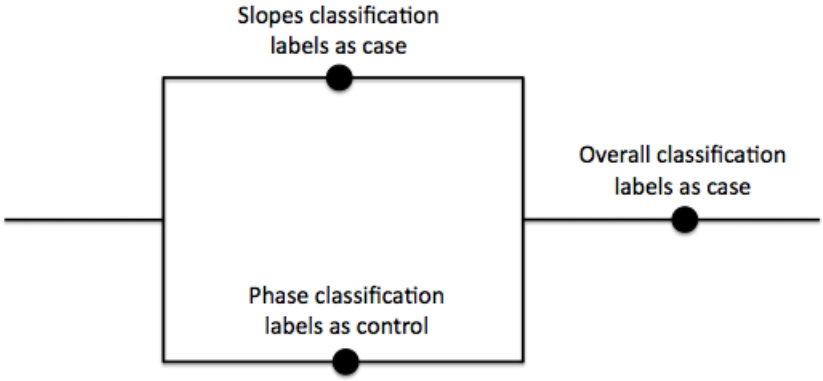


Figure 30: Combining classification procedures: parallel.

3.1.6 Classification Results

Table 3 displays the SVM classification results on the testing set by classification procedure. The best individual classification procedure in terms of overall accuracy rate is the classification based on time-varying slopes (mean accuracy = 64.59%). The best combined classification procedure in terms of overall accuracy rate is the one obtained through parallel combining (mean accuracy = 65.67%). This parallel combined procedure has the additional advantage of a very high sensitivity rate (88.8%), meaning it is highly capable of detecting cancer from the samples. The drawback is the low specificity rate (42.36%), meaning the procedure often identifies samples as cancerous when they are not. Therefore, this type of procedure should not be used

Table 3: Classification results

Classification Procedure	Mean Accuracy Rate	Mean Sensitivity	Mean Specificity
Slopes	0.6459	0.5961	0.6962
Phase	0.6157	0.6452	0.5862
Combined Serial	0.6304	0.3909	0.8739
Combined Parallel	0.6567	0.8880	0.4236

to suggest to individuals that they likely have ovarian cancer, as it would sometimes cause unnecessary psychological pain and suffering. Instead, this type of procedure could be used for preliminary screening to identify which patients should be more closely monitored for the disease. In addition, we suggest this wavelet-based scaling approach may be combined with existing methodologies considering peaks in the mass spectra in order to improve overall classification performance.

3.2 Breast Cancer Diagnostics

3.2.1 Introduction

Breast cancer is the most common form of cancer in females and the second most common cause of cancer-related death for females in the United States (the National Cancer institute estimated 232,000 new cases and 40,000 fatalities for the year 2014). Since early detection can improve a patient's prognosis as well as provide less invasive interventions, mammography is widely used for screenings with the goal of early detection and treatment (National Cancer Institute, 2014). However, mammography has limitations. The radiological interpretation of mammogram images is complicated by the heterogeneous nature of normal breast tissue and the fact that cancers are often of the same radiographic density as normal tissue. As a result, sensitivity may be affected, especially in women with dense breasts. The National Cancer Institute estimates 20% of tumors present at the time of screening are undetected (National Cancer Institute, 2014). Furthermore, researchers have found that, in general, breast tissue is denser among younger women, potentially making it even more difficult to

detect tumors. In a study of over 300,000 screening mammograms, Carney et al. (2003) observes 31% of cancers are undetected in women 40 to 44 years of age as opposed to 17% in women 80 to 89 years of age. Specificity is a concern as well – according to the Lancet, of the 5% of mammograms that suggest further testing, as high as 93% show up as false positives (Houssami et al., 2006).

As a result of these challenges, recent research has investigated the use of computer-aided detection (CAD). In a 2001 study, over 12,000 screening mammograms were interpreted first without the assistance of CAD, then reinterpreted with the suspicious regions marked by the CAD system (Freer and Ulissey, 2001). The authors observed a 19.5% increase in the number of cancers detected and an increase in the proportion of early-stage (0 and I) malignancies detected from 73% to 78%. Most CAD algorithms rely on pattern recognition and attempt to identify physical characteristics of microcalcifications specifically (Chan et al., 1987; Cheng et al., 2003).

More recently, researchers have taken a different approach to identifying breast cancer on mammograms by utilizing the concepts of self-similarity, scaling, and fractality. A 2012 study using the discrete complex wavelet transform on mammogram images obtained a classification procedure based on the spectral slopes and phase variance of mammograms with and without cancer with an accuracy rate of nearly 86% (Nicolis et al., 2012). However, it was later discovered that the mammograms of the cases were performed on a different mammography unit than the mammograms of the controls. Therefore, it is unclear how much of the separation in the data is due to the presence of the cancer and how much is due to the difference in mammography unit. However, the advantage of this general approach is that it captures information contained in the background tissue of images rather than only relying on lesion conspicuity.

In this study, we use wavelets to investigate the spectral slopes of mammograms

with and without cancer, removing the mammography unit effect since all mammograms were obtained from the same unit. We show how these slopes may be used in a classification procedure to build a classifier for separating cases from controls. In addition, we consider two asymmetry statistics in order to form additional features to improve the classification result.

3.2.2 Data

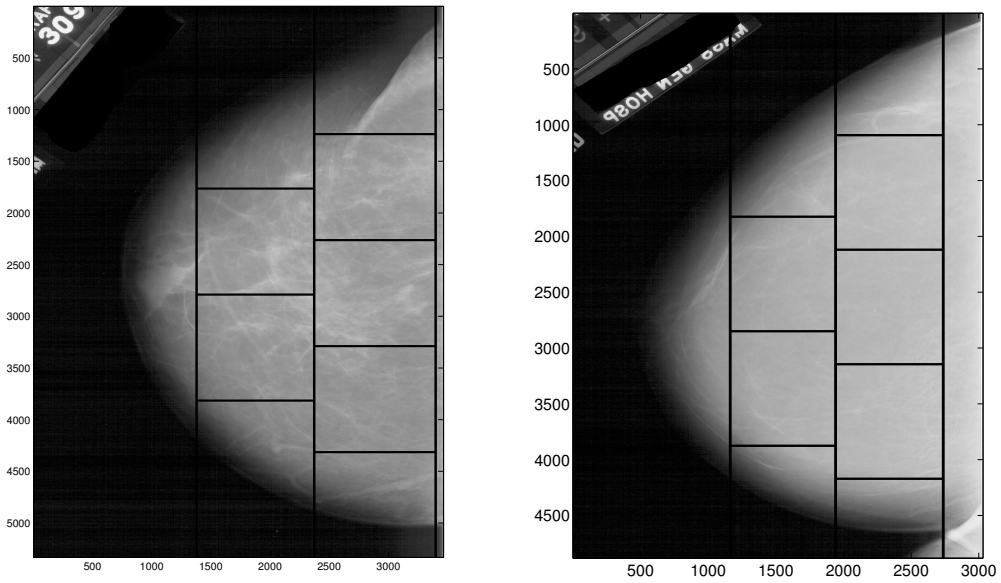


Figure 31: Mammogram images, with cancer (left) and without cancer (right), split into five sub-images each.

A collection of digitized mammograms for analysis was obtained from the University of South Florida's Digital Database for Screening Mammography (DDSM) (Heath et al., 2001, 1998). Images from this database containing suspicious areas are accompanied by pixel-level "ground truth" information relating locations of suspicious regions to what was assessed and verifed through biopsy. This image analysis is based on 79 cases and 45 controls, all scanned on the HOWTEK scanner at the full 43.5 micron per pixel spatial resolution. Each case study contains craniocaudal (CC)

and mediolateral oblique (MLO) projection mammograms from a screening exam. However, only the CC projection mammogram for a single breast is analyzed, the breast containing a cancer for cases and either of the breasts for controls.

Each image is split into five sub-images, each of size 1024 x 1024 pixels. Dividing the images in this way allows the capturing of only breast tissue, and smaller portions of the image data may be analyzed at a time. Figure 31 shows examples of a mammogram with cancer and one without cancer, split into five sub-images each.

3.2.3 The Scale-Mixing Transform and Spectral Slope

The 2-D scale-mixing wavelet transform is applied to regions in both sets of images (cases and controls), and wavelet spectra are formed. The Symmlet 8 tap filter is used, as it provides a compromise between smoothness and locality of support. Figure 32

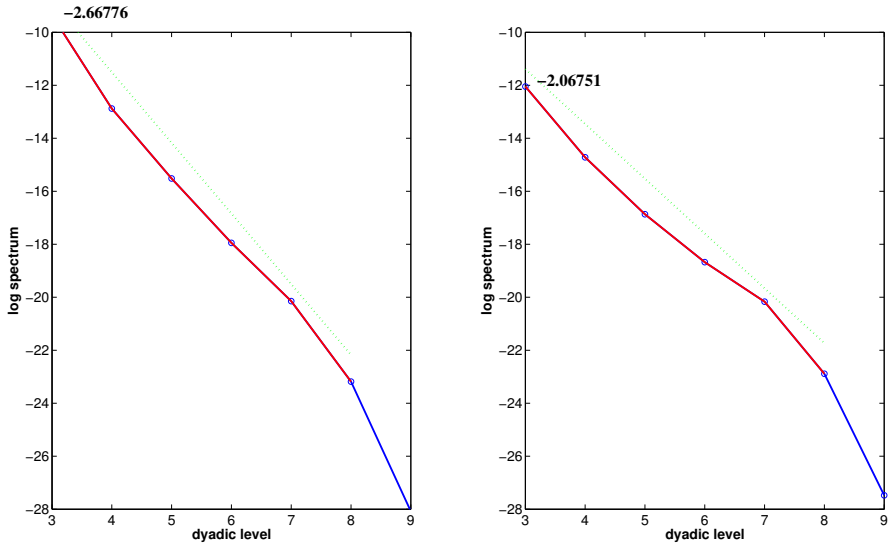


Figure 32: Log energy spectra for a single region of a case (left) and the log energy spectra for the corresponding region of a control (right)

shows the log energy spectra formed for single corresponding regions from the images in Figure 31. The slope of energies in the diagonal hierarchy across various dyadic levels is more negative for the cancerous breast tissue, indicating more regularity. In the one-dimensional case (e.g. time series), high regularity means having long-term

positive autocorrelation. In other words, a high value in the series will probably be followed by another high value, and the values a long time into the future will also tend to be high. This concept of regularity may also be applied to the two-dimensional case (e.g. images), as is illustrated in Figure 33.

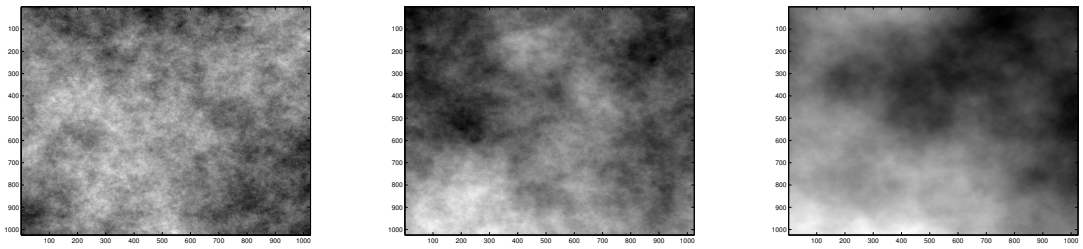


Figure 33: Examples of images with low (left), moderate (middle), and high (right) regularity.

3.2.4 Asymmetry Statistics

In addition to the spectral slope, we consider two statistics representative of differences in energies of mixed detail levels. First, we consider a studentized asymmetry statistic, defined as

$$t_{ij} = \frac{\bar{e_i} - \bar{e_j}}{\sqrt{\sigma^2_{e_i}/n_{e_i} + \sigma^2_{e_j}/n_{e_j}}},$$

where e_i represents energy at fixed dyadic level pairing i above the diagonal and e_j represents energy at fixed dyadic level pairing j below the diagonal where j is the reverse pairing of i. Since wavelets are decorrelating, these sets of energies are approximately independent. Figure 34 demonstrates the wavelet coefficients contributing to this statistic. Note that in regions above the diagonal hierarchy, the level of detail is greater in the horizontal direction than in the vertical direction. In regions below the diagonal hierarchy, the level of detail is greater in the vertical direction than in the horizontal direction.

We also consider a fold change asymmetry statistic, defined as

$$fc_{ij} = \frac{\bar{e_i}}{\bar{e_j}},$$

where e_i represents energy at fixed dyadic level pairing i above the diagonal and e_j represents energy at fixed dyadic level pairing j below the diagonal where j is the reverse pairing of i. Use of an asymmetry statistic of this form is motivated by fold change statistics often used in analyzing similarly structured microarray data Tibshirani (2007).

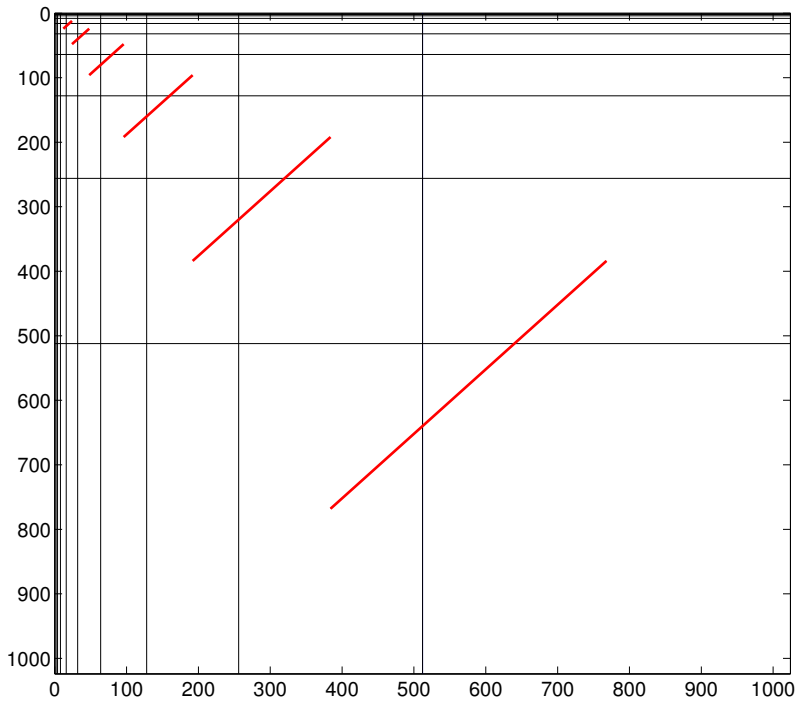


Figure 34: Illustration of wavelet coefficients contributing to the asymmetry statistics. Red lines connect the regions of the wavelet-transformed image used in the calculations.

In order to demonstrate how vertical and horizontal features in an image may be captured through an asymmetry statistic, we compute the asymmetry statistics for two highly directional 256 x 256 pixel images of cardiac tissue obtained from the

University of Delaware's Department of Biological Sciences website (`http://www.udel.edu/biology/Wags/histopage/colorpage/cmu/cmu.htm`). The two images, one of central nuclei and striations having strong vertical features and one of skeletal muscle vasculature having strong horizontal features, are shown in Figure 35. The asymmetry statistics at various dyadic level pairings for these two images are displayed in Table 4. The lower dyadic level pairings represent more coarse features of each image, while the higher dyadic level pairings represent more fine features of each image. The systematic differences in asymmetry statistics may be seen for the coarser level pairings (2 and 3, 3 and 4, 4 and 5). The finer level pairings (5 and 6, 6 and 7) are not capable of "seeing" the images' directional differences. Referring back to Figure 34, the studentized asymmetry statistics for the image with prominent vertical features are positive and the fold change asymmetry statistics are greater than 1, meaning there is more energy in regions above the diagonal hierarchy. In contrast, the studentized asymmetry statistics for the image with prominent horizontal features are negative and the fold change asymmetry statistics are less than 1, meaning there is more energy in regions below the diagonal hierarchy.

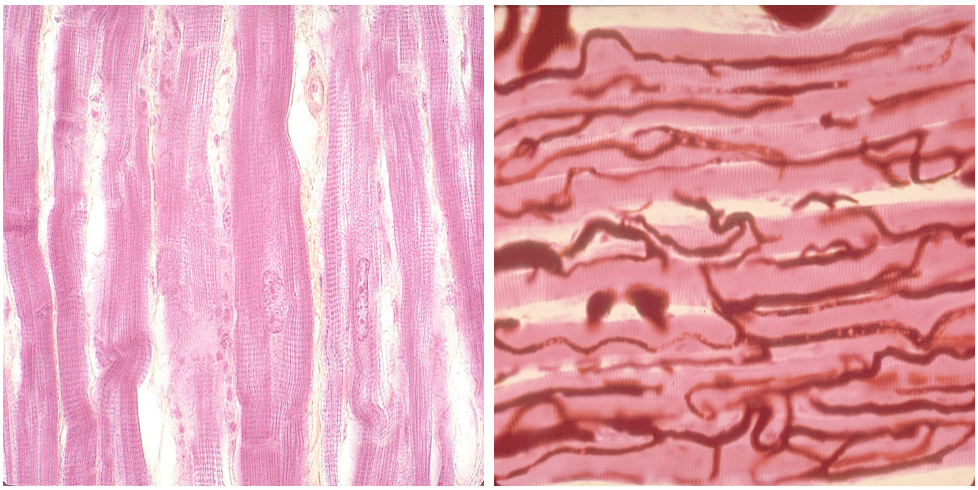


Figure 35: Image with strong vertical features (left) and image with strong horizontal features (right).

In order to calibrate the asymmetry statistics for any image, we may find the

Table 4: Comparison of the asymmetry statistics for images with strong vertical and horizontal features by dyadic level pairing. The systematic differences in asymmetry statistics may be seen for the coarser level pairings (2 and 3, 3 and 4, 4 and 5). The finer level pairings (5 and 6, 6 and 7) are not capable of "seeing" the images' directional differences.

	2 and 3	**3 and 4**	**4 and 5**	**5 and 6**	**6 and 7**
Studentized Asymmetry Statistic					
Vertical	2.5219	2.6747	4.3881	4.4380	0.5748
Horizontal	-1.9690	-4.1924	-3.9102	1.0322	4.7830
Fold Change Asymmetry Statistic					
Vertical	2.9400	2.0928	1.9648	1.3481	1.0251
Horizontal	0.3920	0.3025	0.3284	1.0980	1.1780

degree of deviation from isotropy (uniformity in all orientations) in the following way: We simulate a large number (say 1,000) of fractional Brownian fields (fBfs) with the same spectral slope as the image. Then, we find the empirical bootstrap distributions of each asymmetry statistic. Finally, we evaluate where in the distribution the actual image's asymmetry statistics fall by computing associated achieved significance levels (ASLs). These ASLs are analogous to p-values, with values close to zero indicating significant deviation from isotropy.

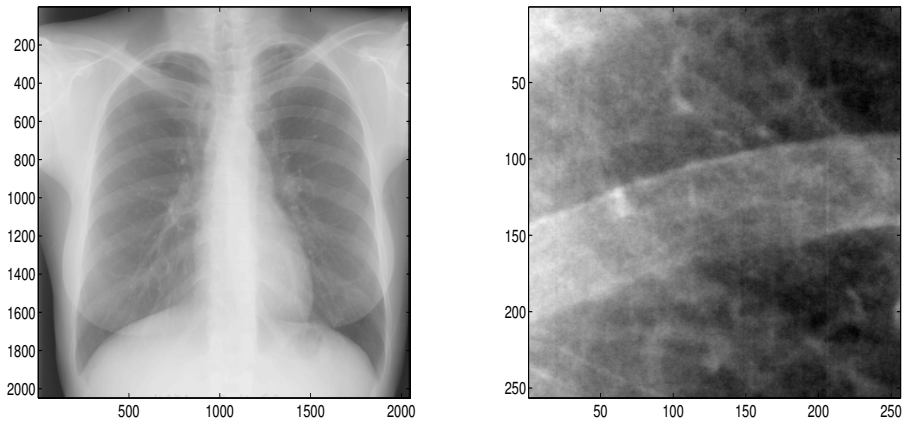


Figure 36: Chest radiograph (left) and the portion analyzed for deviation from isotropy (right)

For example, consider the 256 x 256 portion of the chest radiograph shown in

Figure 36. First, we find the spectral slope (-2.8529) and generate 1,000 fBfs with the same slope. We then compute asymmetry statistics corresponding to each fBf and generate the empirical bootstrap distributions of the asymmetry statistics at the three coarsest levels. Figures 37 and 38 show these bootstrap distributions (histograms) and where the actual image's asymmetry statistics fall in the distributions (red vertical lines). This chest radiograph has asymmetry statistics falling in the left tails of the distributions, indicating high horizontal directionality (t statistic ASLs (coarse to fine): 0.061, 0.001, 0; fc statistic ASLs (coarse to fine): 0.051, 0, 0). Therefore, the asymmetry statistics pick up the coarse horizontal directionality of the rib in the image.

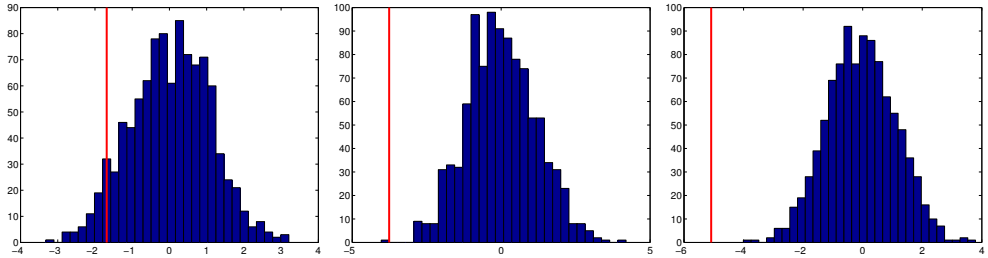


Figure 37: Empirical bootstrap distributions (histograms) for the t statistic at the three coarsest levels and where the chest radiograph subimage's asymmetry statistics fall in the distributions (red vertical lines). This image has asymmetry statistics falling in the left tails of the distributions, indicating high horizontal directionality (t statistic ASLs (coarse to fine): 0.061, 0.001, 0).

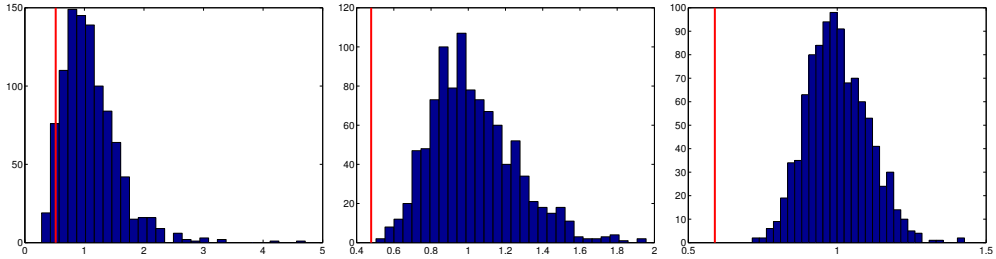


Figure 38: Empirical bootstrap distributions (histograms) for the fc statistic at the three coarsest levels and where the chest radiograph subimage's asymmetry statistics fall in the distributions (red vertical lines). This image has asymmetry statistics falling in the left tails of the distributions, indicating high horizontal directionality (fc statistic ASLs (coarse to fine): 0.051, 0, 0).

3.2.5 Comparing Descriptors for Cases and Controls

As mentioned in section 3.2.3, the spectral slope seems to be more negative for the mammograms with cancer than for the mammograms without cancer. In order to investigate this relationship analytically, we perform a two-way nested ANOVA, under the model

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}, \epsilon_{ijk} \sim N(0, \sigma^2),$$

where y_{ijk} represents spectral slopes for each possible region of each mammogram, $\alpha_i, i = 1, 2$ represents the effect of case/control on the slope, $\beta_{j(1)}, j = 1, ..., 79$ represents the effect of the person on the slope for cases, and $\beta_{j(2)}, j = 1, ..., 45$ represents the effect of the person on the slope for controls. This analysis separately models the effects of the person and whether the mammogram contains cancer on the spectral slope. Table 5 contains the two-way nested ANOVA results, showing that cases and controls produce significantly different spectral slopes ($\hat{\alpha}_1 = -0.0462$, corresponding to cases, $\hat{\alpha}_2 = 0.0462$, corresponding to controls). There are also significant differences in spectral slopes among people. For our classification procedure, we use slopes $y_{ijk} - \hat{\epsilon}_{ijk}$ to represent each image.

Table 5: Two-way nested ANOVA on spectral slopes

Source	SS	df	MS	F	Prob>F
Case/Control	1.2251	1	1.22511	8.39	0.0045
Person(Case/Control)	17.821	122	0.14607	6.53	0
Error	11.0909	496	0.02236		
Total	30.137	619			

We fit analogous two-way nested ANOVA models for values of both the studentized asymmetry statistic and the fold change asymmetry statistic at each detail level. For both asymmetry statistics, cases and controls produce significantly different values at all but the coarsest levels of detail. Both t and fc statistics tend to be higher for cases

than controls. The person effect is significant for each asymmetry statistic at every detail level. Tables 6 and 7 contain the two-way nested ANOVA results for values of the t and fc statistics, respectively, at dyadic level pairing 5 and 6 (as an example). For this level pairing, the fitted t statistic coefficients are $\hat{\alpha}_1^* = 0.4660$ for cases and $\hat{\alpha}_2^* = -0.4660$ for controls. The fitted fc statistic coefficients are $\hat{\alpha}_1^{**} = 0.0219$ for cases and $\hat{\alpha}_2^{**} = -0.0219$ for controls. As with slopes, we use asymmetry statistic values with subtracted fitted residuals to represent each image.

Table 6: Two-way nested ANOVA on t statistics

Source	SS	df	MS	F	Prob>F
Case/Control	124.52	1	124.52	5.4	0.0218
Person(Case/Control)	2814.36	122	23.069	3.1	0
Error	3696.7	496	7.453		
Total	6635.58	619			

Table 7: Two-way nested ANOVA on fc statistics

Source	SS	df	MS	F	Prob>F
Case/Control	0.2739	1	0.2739	7.24	0.0081
Person(Case/Control)	4.6145	122	0.03782	3.05	0
Error	6.1441	496	0.01239		
Total	11.0325	619			

3.2.6 Classification Results

Table 8 displays the support vector machines (SVM) classification results by choice of kernel (linear, quadratic, or radial basis) with and without asymmetry statistics for 1,000 iterations. For each iteration, the data set is split into a 70% training set (87 images) and a 30% testing set (37 images). For each procedure, the spectral slopes are used as features to train the model. The addition of asymmetry statistics of either form as features in the classification significantly raises the sensitivity and

overall accuracy. The best results using only spectral slopes in the classification are obtained using the linear kernel (accuracy = 62.68%). The best overall results are achieved using the linear kernel and including both spectral slopes and fold change asymmetry statistics in the classification (accuracy = 76.59%).

Table 8: SVM classification results. The best results are achieved using the linear kernel and including both spectral slopes and fold change asymmetry statistics in the classification.

	Mean Accuracy Rate	Mean Sensitivity	Mean Specificity
Slopes Only			
Linear	0.6268	0.7039	0.4952
Quadratic	0.6027	0.6855	0.4607
Radial Basis	0.6017	0.5978	0.6099
Slopes + Asymmetry t			
Linear	0.6731	0.6685	0.6819
Quadratic	0.6684	0.6930	0.6314
Radial Basis	0.6332	0.6901	0.5474
Slopes + Asymmetry fc			
Linear	0.7659	0.7250	0.8379
Quadratic	0.7195	0.7414	0.6856
Radial Basis	0.7296	0.7800	0.6513

3.2.7 Conclusion

Mammography is routinely used to screen for breast cancer. However, the interpretation of mammograms by radiologists is made difficult by the heterogeneous nature of normal breast tissue and the fact that cancers are often of the same radiographic density as normal tissue. CAD algorithms have been developed to assist in the identification of suspicious regions. However, most CAD algorithms rely on pattern recognition and attempt to identify physical characteristics of microcalcifications specifically. By using scaling properties of mammogram images, we have additionally captured information contained in the background tissue of images which is not utilized when only considering lesion morphologic. Using features based on spectral slopes and our defined asymmetry statistics, we have achieved a SVM classification procedure with 76.59% accuracy on the testing sample. Importantly, the mammograms of both the

cases and controls were performed on the same mammography unit, so this level of separation is not due to any image acquisition effect. We suggest that this classifier may be used in conjunction with other methodologies in order to improve the detection of breast cancer through mammography. If the tool proves robust on further investigation, it may be useful in outlining cases that require heightened scrutiny or even addition of supplemental screening modalities, especially in cases where the patient has dense background parenchymal tissue, a factor known to decrease mammographic sensitivity.

3.3 Lung Cancer Diagnostics

3.3.1 Introduction

Lung cancer is the second most common cause of cancer in both males and females, and the most common cause of cancer related death. The American Cancer Society estimates that this year there will be more than 226,000 newly diagnosed cases of lung cancer in the United States, with number of deaths exceeding 160,000. The mortality of lung cancer patients is related to the stage at which cancer is diagnosed, with improved survival for cancers detected earlier (Goldstraw and Crowley, 2007; Aberle DR et al., 2011).

Chest radiography (CXR) is one of the most commonly performed radiologic examinations worldwide, and lung cancer is often diagnosed on chest radiography. Consequently, it would be helpful to be able to detect nodules on chest radiography with high sensitivity. The sensitivity of chest radiography for detection of pulmonary nodules 7-20 mm is approximately 46% (Oda and et all., 2010). A recent study showed that 75% of perihilar and 90% of peripheral nodules that are identifiable on chest radiography were missed at the time of initial interpretation (Meziane and et al., 2012). Nodules not detected early on radiography may be detected later, perhaps once the patient has become symptomatic due to late stage disease. Consequently,

efforts have been made to develop software for computer aided detection (CAD) of small pulmonary nodules.

Several CAD systems have been reported for detecting lung nodules on CXR images usually adopting pattern recognition approach, which includes a feature extraction and statistical classification. Nodules show up in a CXR image as ovoid intermediate attenuation lesions. Several CAD schemes start with enhancing nodules by filtering with nodule-like filters and suppressing the background. Also, preprocessing techniques are applied such as unsharp masking or similar operations. Nodule candidates are also detected by using template matching or by applying the Hough transform, to list just a few approaches (Klik et al., 2006). One shortcoming of existing CAD approaches is that they make a priori assumptions about nodule shape and, therefore, may fail to identify atypical nodules. The approach taken in this work is to perform an image texture analysis using wavelet-based scaling.

3.3.2 Data

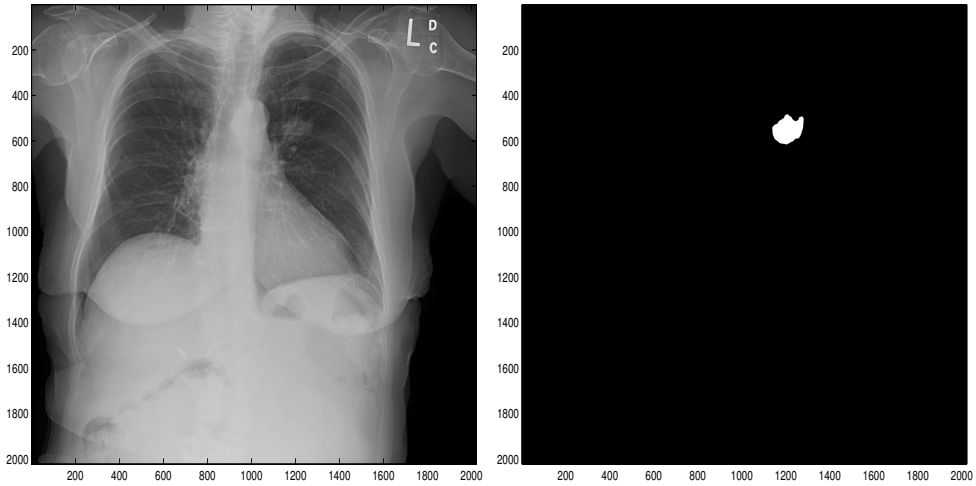


Figure 39: Lung CXR image (left) and mask showing pulmonary nodule location (right)

The data obtained are 21 lung CXR images with pulmonary nodules for which the locations are known from the Lung Image Database Consortium (LIDC). Figure

39 shows one such image, along with a mask providing the nodule location.

Based on the nodule location, we define the region of interest (ROI) as a 256 x 256 pixel region of the image containing the nodule. The region of control (ROC) is then defined as an image the same size as the ROI which is symmetric about the spine (Figure 40).

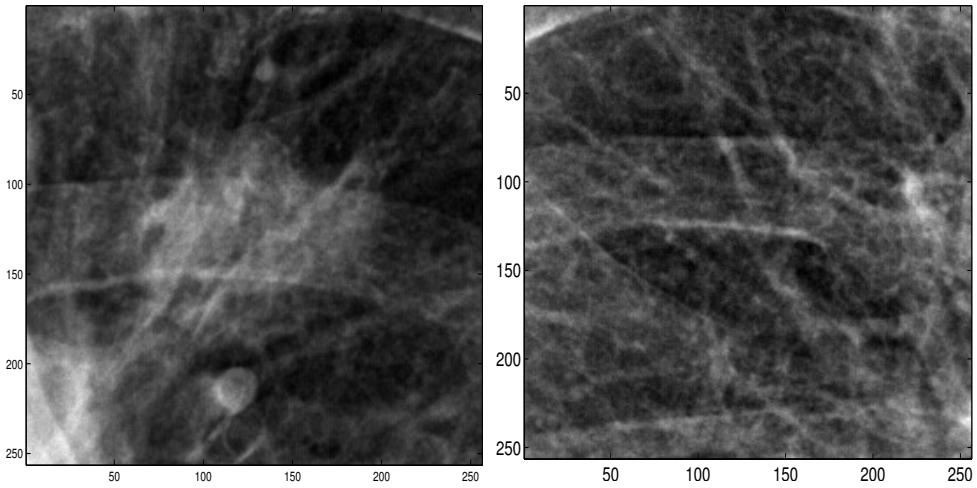


Figure 40: ROI (left) and ROC (right)

3.3.3 The Scale-Mixing Transform, Spectral Slope, and Asymmetry Statistic

The 2-D scale-mixing wavelet transform is applied to both the ROI and ROC, and wavelet spectra are formed. Figure 41 shows the log energy spectra formed for the lung image in Figure 39. In this case, the slope of energies in the diagonal hierarchy across various dyadic levels is more negative for the ROI, indicating more regularity. However, this trend does not hold for all images in the sample. As in the mammography analysis, asymmetry statistics are computed as well.

3.3.4 Classification Results

Based on slopes and asymmetry statistics, classification is performed via support vector machines (SVM) to predict whether the considered lung images contain nodules. As in the mammography analysis, 1,000 iterations of this classification procedure are

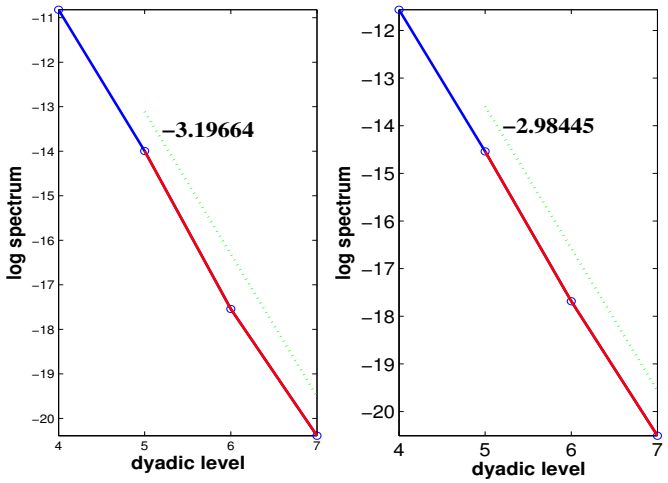


Figure 41: Log energy spectra for ROI (left) and ROC (right)

performed. In each iteration, 15 of the 21 lung images are randomly selected as a training set, and the remaining 6 become the testing set. The results are displayed in Table 9. Interestingly, use of the quadratic kernel yields the best results for the lung CXR classification (accuracy = 58.79%), whereas use of the linear kernel yielded the best results for mammogram classification. Also, in contrast to mammogram results, using the studentized asymmetry statistics yields higher accuracy rates than using the fold change asymmetry statistics.

Table 9: SVM classification results - lung CXRs

	Mean Accuracy Rate	Mean Sensitivity	Mean Specificity
Slopes + Asymmetry t			
Linear	0.4576	0.4617	0.4535
Quadratic	0.5879	0.6523	0.5235
Radial Basis	0.4797	0.4927	0.4667
Slopes + Asymmetry fc			
Linear	0.5077	0.5927	0.4227
Quadratic	0.5383	0.6010	0.4757
Radial Basis	0.4485	0.4312	0.4658

Our accuracy in classifying lung images is significantly lower than our accuracy in classifying mammograms. One possible explanation is that the sample of lung CXRs is quite small (less than half the size of the mammography sample). An alternative

explanation is that this performance difference may be due to heterogeneous anatomic structures in three dimensions being represented two dimensionally in the lung images. Lung CXRs include not only the lung tissue, but bone tissue and potentially tissues of the heart (depending on the ROI location). Such other anatomic structures behind or in front of the lung tissue may mask key differences in the nature of the lung tissue in the area of a nodule. Mammograms, on the other hand, are much more homogeneous in terms of anatomic structure, potentially leading to better classification.

CHAPTER IV

ASSESSING THE IMPACT OF SOCIAL MEDIA DATA IN COMMERCIAL CREDIT MODELING VIA RANDOM FORESTS

4.1 Introduction

It comes as no surprise that an individual's past behavior is strongly predictive of future behavior. Traditional credit risk models leverage this fact by incorporating independent variables such as the number of times an individual has been late on a credit card payment in the last twelve months in order to predict whether s/he will be late on a payment in the next twelve months. Lenders subsequently use such predictions, generally referred to as scores, to make decisions pertaining to lending and account management.

A similar notion extends from risk management purposes to those related to prospecting and lead generation. For example, an individual might tend to make balance transfers in a cyclical manner, causing a marketing team to direct relevant offers to that individual in advance of the anticipated transaction. Similarly, suppose within the commercial lending space a small business is observed to have grown steadily in terms of factors such as number of customers, number of employees, or annual revenue. These might be indicators that the business is a likely candidate to soon apply for a new loan in order to expand, again enabling the marketing team to take appropriate action. These types of predictions are commonly made using response or propensity models.

4.1.1 Typical Scoring Methodology

Most often, when institutions use an individual's credit score for risk-decisioning purposes, they must provide to the individual reasons for any adverse action(s) taken, such as rejection of a loan application or reduction of an existing credit limit. For this reason, logistic regression is the most common methodology used to develop credit scoring models, as the extraction of so-called "reason codes" from such models is generally quite straightforward. Another commonly used technique, often for segmentation, is CART. To a lesser extent, more complex algorithms such as spline approaches and neural networks are implemented. Mays (2001) and Finlay (2010) provide thorough overviews of the most commonly used modeling methodology.

On the other hand, models geared more towards marketing than risk need not be capable of generating reason codes. For example, if a bank chooses not to extend a pre-screen offer to an individual, they need not disclose to that individual why they elected not to do so. For this reason, a broader class of modeling approaches might be applied in the marketing domain, including machine learning approaches such as random forests and support vector machines. Hastie et al. (2001) provides a detailed overview of these, and many other machine learning algorithms.

4.1.2 The Emergence of Social Media Data in Decision-Making

In the last decade, social media data has taken a prevalent role in assisting institutions with various forms of decision-making. Such data has a wide range of applicability, ranging from decisions pertaining to lending and marketing, as discussed in more detail below, to those associated with employment, such as hiring decisions, as discussed in Roth et al. (2013) and Brown and Vaughn (2011).

A plethora of institutions, including large players such as Lending Club and On-Deck, participate in the marketplace lending industry, also known as peer-to-peer

lending. A marketplace platform enables otherwise idle lenders and borrowers to locate each other and come to the terms of a loan, all without the involvement of banks and card issuers. Magee (2015) estimates the overall size of the marketplace to be approximately one trillion dollars.

Marketplace lending typically differs from traditional lending, both with respect to the type of data on which decisions are based, as well as the modeling methodology. For example, Upstart focuses on finding high-quality consumer borrowers who lack long credit and employment histories. This group consists primarily of young individuals, who happen to be those most likely to have profiles available on professional networking websites such as Linkedin. Upstart will begin with a traditional credit score such as FICO (if available), and then augment the score with social media data elements such as university attended, major, GPA, and job-title. Other factors obtained from social media might also be taken into account. As described in Armour (2014), individuals with many friends, none of whom are close friends, might be perceived as having a different risk-profile than those with a small number of friends, all of whom are close.

The goals of models that incorporate these new data sources are the same as those of any credit scoring model, namely to help a lender to decide whether or not to lend, and to assist the lender with establishing loan terms, such as size and interest rate. Modeling methodology might be as simple as a segmentation or regression approach, or it could be based on more complex, machine learning algorithms.

While viable on the surface, a framework based on such user-controlled information might easily introduce various complexities. If consumers have an idea how the decisions are made, they have an opportunity to deliberately try to increase their scores. This could be done by falsifying information on a Linkedin profile, or as examined in Wei et al. (2014), establishing a friend network that will be interpreted most favorably by the scoring models. Rusli (2013) elaborates these points, including

a discussion on FICO's stance that lending decisions should not be based on how many Facebook friends an applicant has. Lin et al. (2013) also investigates problems associated with information asymmetry. Dixon and Gelman (2014) discusses how new scores, based on thousands of factors unknown to those being scored and geared towards a host of applications, including risk, fraud, health, and insurance, potentially threaten consumers with respect to a variety of factors, including privacy, fairness, and due process. Lohr (2015) describes how relying on complicated machine learning algorithms could easily lead to cases of discrimination, and indicates that while the landscape is not currently well-established, regulators such as the Consumer Financial Protection Bureau, while encouraging of innovation, are closely monitoring the ongoing evolution of the industry.

Perhaps of a lesser concern from a discriminatory and regulatory point of view, and also relevant to the ensuing discussion, is the utilization of social media for marketing purposes. Goel and Goldstein (2013) finds social networks highly capable of determining individual behavior as it pertains to actions such as patronizing a brick-and-mortar department store and joining recreational leagues. Furthermore, they demonstrate a significant benefit from adding social media data into traditional models. Hill and Volinsky (2006) uses a variety of statistical methodology to investigate similar phenomena.

4.2 Background and Considerations

Bureau data and the models they generate can be classified according to a variety of criteria. A typical example is vertical, different examples of which include mortgage, auto, credit card, and non-financial accounts such as utilities. An auto-specific scoring model might be used by a lender to evaluate the credit-worthiness of someone applying for a new car loan, while a telco score might be used for someone signing up for a new cell phone contract. Such vertical models are characterized by the dependent

variable on which they are built.

The dependent variable in an auto specific score would typically be obtained by selecting a modeling population of individuals who, at a certain observation date, have an outstanding auto loan. Each loan is tracked for a specified period, which we refer to as the performance window. The value of the dependent variable for a given individual might be set to "bad" if the individual does not conform to the re-payment terms of the loan throughout the performance window. This could occur, for example, if a payment is not made in-full within 60 days of the due date. If all re-payment terms are satisfied, the value of the dependent variable would be set to "good". In some cases, if sufficient information is not available to gauge performance of the borrower or if the borrower's performance falls within a region close to the boundary of the definitions of "good" and "bad", then the dependent variable might be set to "indeterminate". In this case, such an observation would generally be excluded from the model development process.

It is important to note that the independent variables on which vertical models are built need not be vertical specific. For example, the auto-specific score could be built on variables unrelated to the auto vertical such as the existing balance, if any, on an applicant's mortgage account(s).

Bureau data can also be broadly classified according to whether or not an applicant or account holder is identified as a consumer or a business. Some differences in the underlying methodology exist for commercial model development as compared to consumer models. For example, in so-called blended models, data for both a business and the corresponding business-owner might be used for model development. Apart from this and other distinctions outside the scope of this chapter, business behavioral models serve many of the same purposes as consumer models, including the evaluation of traits such as risk and propensity. The remainder of this chapter focuses on developing commercial models for these purposes.

4.2.1 Data Regulations and Restrictions

While many different types of bureau data are likely to be available, usage restrictions and regulations govern the types of data that can be used throughout the model development process. For example, the Commission (1998) (ECOA) makes it illegal for any lender to directly or indirectly discriminate against any of several protected classes, which include race, color, religion, national origin, sex, marital status, and age, unless doing so meets a legitimate business need and is otherwise unavoidable. For this reason, any variable pertaining to any of these classes must generally be excluded from a model designed to make lending decisions. ECOA regulations generally extend from decisions pertaining to consumer lending to those pertaining to commercial lending.

Meanwhile, models designed for marketing purposes are generally subject to a much less rigorous set of regulations than those used to make lending decisions. As we are investigating a type of data not sufficiently vetted for conformity to ECOA guidelines, namely social media data, we have chosen to focus only on the development of models geared towards marketing use-cases.

In addition to the above regulations, model development must also conform to usage restrictions designated by contributors of the data. The restriction relevant to this work relates to the distinction between commercial financial accounts and non-financial accounts. Contributors of the data establish a restriction that financial account records must not be used to develop any model intended to be used for any type of marketing decisioning. Consequently, the models we develop will be commercial marketing models, geared towards lead generation in the non-financial space. The components of lead generation we investigate relate to pre-screening from a risk perspective, and trying to measure the likelihood of a business opening a new non-financial account within a designated period of time. For our purposes, non-financial account shall refer to any account classified as telco, utility, or industrial.

4.2.2 Model Requirements and Methodology

As mentioned above, any model, the outcome of which can lead to a so-called adverse action, must also be capable of generating reason codes for that action. Regulations are well-established with respect to lending, and to a large extent, this limits the statistician's toolset when trying to choose the *best* modeling methodology. It is quite often the case that an algorithm such as random forests or neural networks are not utilized because of the difficulty of reason code extraction from the resulting models, in spite of the fact that such models often significantly outperform corresponding models built using linear or logistic regression.

Regulation is less restrictive in the marketing domain. Consider a marketing model that is used to pre-screen 10 million individuals, for the purpose of trying to identify the 2 million individuals most likely to respond to a given offer. A company that uses a model for this purpose is not required to provide any type of explanation to the 8 million individuals who do not receive the offer. As the models we develop are designed for marketing purposes, they need not be capable of generating reason codes, allowing us to pursue what we feel is the best possible modeling methodology. For this reason, we conduct all analysis using random forest models, a technique originally proposed in Breiman (2001). Throughout the course of the effort, we did consider alternatives, including neural networks and support vector machines, but in our framework, random forests consistently outperformed the other approaches.

4.3 Data Collection

We begin this section with a somewhat peripheral yet interesting discussion on social media data collection. An abundance of literature examines various applications of consumer-level social media data in a variety of domains, some of which we discuss above. Meanwhile, no existing study attempts to measure the ability (or lack thereof)

of commercial-level social media data to function in a predictive role in any of various types of widely implemented behavioral models. In some ways, we found this surprising, given the fact that many popular e-commerce and tourism sites already have done a lot of the work in terms of data aggregation, and in doing so provide an abundance of information about an individual business. Furthermore, one tends to encounter fewer problems associated with non-unique or inaccurate identifiers when examining businesses as opposed to consumers. However, we hypothesize that the lack of existing literature in this commercial domain may stem from researchers not having access to the data necessary to accurately represent businesses' performance.

4.3.1 Collecting Online Hotel Reviews

Businesses falling into the service sector, by far, tend to be those for which online review data is most abundant. Of that group, hotel and restaurant reviews are very common. We initially considered doing a combined study using both hotels and restaurants, but ultimately decided to focus only on hotels, as our initial pilot findings across the two groups were nearly identical.

Many of the hotel review sites aggregate not only their own reviews, but also provide those left by consumers on any of their partner sites. When visiting a hotel page on a tourism site, one commonly finds information pertaining to the hotel such as name and address, a star-rating for the hotel, some high-level summaries of all of the customer reviews that have been registered for that hotel, as well as the text, date, and sometimes username associated with the individual reviews themselves.

Certain sites also have pages which index in some structured manner all of the hotels in their database. The site from which we collected our data uses what they consider to be a 'level-3 index', which at the time of the study, contained 1,910 individual pages. Each of these pages contains an alphabetized list of 100 site-level links for hotels around the world. As we were focusing only on U.S. hotels, our

strategy was to generate a list of the 1,910 index-level URLs, and loop though each, extracting links to any U.S. based hotels. To do so, we used the `R` programming language and its XML package. XML::getHTMLLinks() was applied to each of the index-level pages. Subsequently, grep() was applied to the list with a "united-states-of-america" parameter to subset only on a list of 43,434 U.S. hotels.

With this list of URLs established, we wrote a hotel-level scraper, also in `R`, to extract information for each hotel we wanted to use in the study. The hotel-level scraper relied mostly on XPath syntax to extract information out of the desired node within a page's HTML tree. For example, in any of our list of 43,434 site-level pages, //div[@id='property-name']/div[1]/h1 always returns the name of the hotel. Essentially this XPath statement looks anywhere in the document for a div element with an id attribute having a value equal to property-name, and then extracts the value of the first div element under any such div element. Constructing the XPath for all other information proceeds in a similar manner and shall not be a further focus of this chapter.

4.3.2 Hotel Attribute Creation

In addition to an overall rating out of five stars, the website from which we collected our data provides more specialized ratings for each hotel in five areas: cleanliness, service, comfort, condition, and neighborhood. The reviews consist of the title and body of the text review. The social media data incorporated into our models includes both rating and review information from the website.

Rating data attributes in the models include overall ratings and each of the five specialized ratings listed above as of the observation point in September 2013 (the beginning of the performance window) and as of three months prior in June 2013. In addition, we create attributes which represent the change in rating from June to September by simply taking differences. These rating change attributes are included

in order to capture recent trends which may be predictive of a hotel's likelihood of default or its level of activity. For example, a hotel experiencing financial hardship may need to reduce its staff size, causing a gradual drop in its service rating. In this case, the rating change may be predictive of this hotel defaulting on a loan in the near future. Additional attributes based on ratings include the number of ratings as of both June and September 2013, and the change in number of ratings over the three month time span, again due to its potential connection to a hotel's level of activity.

We also obtain review attributes through textual analysis of the review title. Attributes based on review titles are defined as percentages of a hotel's most recent reviews containing certain words in the title. For example, one attribute would designate the percentage of a hotel's most recent review titles containing the word "comfortable". The words utilized in the modeling (listed below) were chosen from the 150 most frequent review title words overall. We exclude articles (a, an, the) and similar words presupposed to have no discriminatory power.

nice	good	great	price	quiet	excellent
value	clean	staff	service	easy	comfortable
perfect	best	bad	friendly	deal	awesome
ok	need	needs	don't	money	wonderful
stay!	family	airport	convenient		

Only the most recent reviews are incorporated into the review data attributes since these reviews represent the most current customer feedback, and potential customers are likely to only look at the first page or so of reviews on the website. We define a hotel's most recent reviews as the 10 reviews occurring before and closest to the observation point. If a hotel has less than 10 reviews during the observation period, then all review titles are incorporated.

4.3.3 Merging to Bureau Data

The hotel attribute set was merged with hotel information in the database of a major financial information services provider for the set observation date. The online review data alone contained 43,434 hotels. After limiting reviews to those before September 30, 2013, the data set contained information on 36,272 hotels. After matching to the bureau universe of businesses who had non-financial accounts on record, the number of hotels in the data set was reduced to 11,802. Finally, 10,151 had sufficient information available to construct the binary outcomes described in the following section. This modeling sample was then split into a 70% training set and a 30% testing set, containing 7,106 and 3,045 observations, respectively.

There are 453 total attributes in the modeling data set, 50 social media attributes and 403 bureau attributes. Some bureau attributes are based on business information such as geographic region, number of employees, annual sales, and years in business. Other attributes represent specific information about non-financial trade lines, such as the number and age of accounts, balances, utilization, delinquency status, and past due amounts.

One challenging characteristic of the data set is the number of missing values, as roughly 1/4 of the total values are missing. Another challenge is the hundreds of attributes available for modeling. However, the random forest modeling methodology allows the inclusion of missing values as well as a large number of predictor variables without the issues accompanying collinearity in regression methodologies.

4.3.4 Dependent Variables

The two binary dependent variables represent

(1) whether or not a hotel becomes 90 or more days delinquent on any non-financial account, files bankruptcy, or has a charge-off on any non-financial account within 6 months following the observation period, and

(2) whether or not a hotel opened a new non-financial account within 3 months following the observation period.

The activity time window was chosen to be shorter than the default time window so that the model could generate leads most likely to need credit in the immediate future. As a practical example, if a marketing team launches a mail campaign, those who respond are typically more likely to do so shortly after exposure to the promotion rather than several months later. The default time window was chosen to allow enough time to elapse beyond new account openings so that businesses would have enough time to default and records of those defaults would have enough time to be reported.

4.4 Modeling Methodology and Results

4.4.1 Random Forest Overview

The random forest procedure, combining the existing concepts of bagging and random split selection, was developed by Leo Breiman in 2001 (Breiman, 2001). Breiman defines a random forest as a classifier consisting of a collection of tree-based classifiers

$$h(\mathbf{x}, \Theta_k), k = 1, 2, ..., K$$

where Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$.

The procedure for building a random forest model is as follows: sample the data set with replacement n times, where n is the number of trees in the forest model, specify the number of variables randomly selected at each tree node to be considered for splitting, grow each tree to the fullest extent, and prune each tree if desired. Figure 42 shows an example of a binary decision tree which may be included in a random forest model. At node A, the entire bootstrap sample of the data is present. This data set is then split into nodes B and C, based on the value of a single discriminatory variable. The process continues down the tree until leaf nodes D, H, I, F, and G are reached.

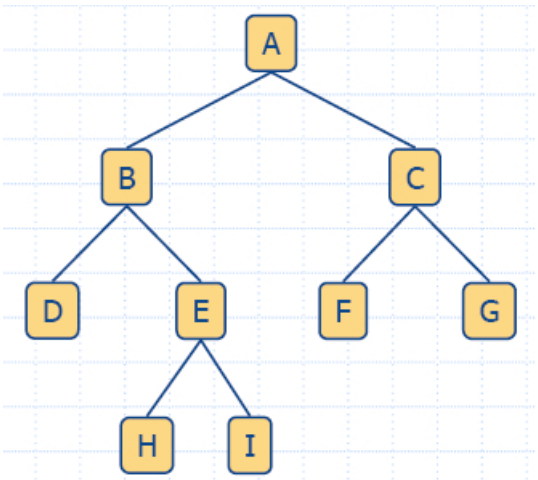


Figure 42: Binary decision tree

Breiman highlights several advantages to random forest models. One advantage is the existence of an upper bound on the generalization error. Generalization error is defined in terms of the margin function, $mr(\mathbf{X}, Y)$, measuring the extent to which the average number of votes for the correct class exceeds the average number of votes for any other class. The margin function for a random forest is

$$mr(\mathbf{X}, Y) = P_{\Theta}(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_{\Theta}(h(\mathbf{X}, \Theta) = j)$$

and the strength of a set of classifiers $h(\mathbf{x}, \Theta)$ is

$$s = E_{\mathbf{X},Y} mr(\mathbf{X}, Y)$$

Classifiers with higher strength predict the correct class more often. The generalization error is defined as

$$PE^* = P_{\mathbf{X},Y}(mr(\mathbf{X}, Y) < 0),$$

where the subscripts $\mathbf{X}, Y$ indicate the probability is over the $\mathbf{X}, Y$ space.

In order to find the upper bound on generalization error, Breiman defines the raw margin function as

$$rmg(\Theta, \mathbf{X}, Y) = I(h(\mathbf{X}, \Theta) = Y) - I(h(\mathbf{X}, \Theta) = \hat{j}(\mathbf{X}, Y)),$$

where $I()$ is the indicator function and $\hat{j}(\mathbf{X}, Y)$ is the most likely class vote other than the correct one. Note that $mr(\mathbf{X}, Y)$ is the expectation of $rmg(\Theta, \mathbf{X}, Y)$ with respect to Θ. Let $\rho(\Theta, \Theta')$ denote the correlation between $rmg(\Theta, \mathbf{X}, Y)$ and $rmg(\Theta', \mathbf{X}, Y)$ where Θ and Θ' are fixed values of independent identically distributed random variables. Let $sd(\Theta)$ denote the standard deviation of $rmg(\Theta, \mathbf{X}, Y)$ with Θ fixed. Then

$$PE^* \leq \frac{\bar{\rho}(1 - s^2)}{s^2},$$

where

$$\bar{\rho} = \frac{E_{\Theta,\Theta'}(\rho(\Theta, \Theta')sd(\Theta)sd(\Theta'))}{E_{\Theta,\Theta'}(sd(\Theta)sd(\Theta'))}$$

This upper bound clearly reflects the two ingredients involved in the generalization error for random forests: the strength of the individual classifiers in the forest, and the correlation between them in terms of the raw margin functions.

Other advantages of the random forest method are its ability to handle thousands of variables without deletion, to handle missing values, to detect variable interactions, and to represent nonlinear relationships.

Although Breiman's random forest methodology has only been around since 2001, there is already increasing evidence in the literature that random forests can achieve better predictive accuracy than logistic regression in modeling credit data, particularly when multicollinearity is present and there are complex interrelationships among the predictors (Sharma, 2012; Kruppa et al., 2013). In a 2012 study, logistic regression and random forest model performance are compared on a popular SAS data set pertaining to home equity loans and on proprietary credit data (Sharma, 2012). For the home equity data, the logistic regression model achieved an area under the curve (AUC) of 0.78, while the random forest model achieved an AUC of 0.92. The random forest model also produced significant lift over the logistic regression model when fitted on the much larger credit data set (logistic regression model AUC=0.70, random forest model AUC=0.85).

In a study of the likelihood of consumers to default on payment of installment purchases of household appliances, Kruppa et al. (2013) demonstrate the superior performance of random forest probability estimation trees to not only logistic regression, but also to k-nearest neighbors and bagged k-nearest neighbors. Ching-Chiang et al. (2012) utilize, in particular, random forests' identification of variable importances in their modeling. They use variable importances from random forests to select variables for use in decision tree, classification and regression trees (CART), support vector machine (SVM), and rough set theory (RST) models. By using this modeling strategy and incorporating market-based information, they obtain accuracy rates as high as 93.4% in prediction relating to commercial credit ratings.

4.4.2 Our Modeling Approach

Four separate random forest models are built:

(1) a default model using only social media data as inputs,

(2) a default model using only traditional bureau data as inputs,

(3) an activity model using only social media data as inputs, and

(4) an activity model using only traditional bureau data as inputs.

Along with comparing these individual models directly, we wish to compare each individual model to a combined model with all attributes. However, we do not simply build a random forest model with all 453 attributes directly. Since there are nearly 10 times more traditional bureau attributes than social media data attributes, a combined random forest model of this nature may be heavily dominated by this more traditional data and not experience any lift from the addition of social media data. Instead, we use the proportions of class votes from each pair of individual random forest models to build two logistic regression models with main effects and the first order interaction term, one to predict the overall likelihood of default on a

non-financial account and another to predict the overall likelihood of opening a new account. In each case, we fit the model

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

$$P(Y = 1) = p,\ P(Y = 0) = 1 - p$$

where p is either the probability of default or the probability of opening a new account, x_1 is the proportion of class 0 votes from the social media model, x_2 is the proportion of class 0 votes from the bureau model, and Y is the binary outcome.

4.4.3 Tuning Parameters

4.4.3.1 Splitting Criterion

Several impurity measures may be used to decide how to split at each node of the classification trees. One possible choice is entropy,

$$E = -\sum_i p_i \log_2 p_i,$$

where i refers to the class. To decide on the best possible split, compute the entropy before splitting, compute the entropy resulting from each possible split, and choose the split with the greatest decrease in entropy. Then the classification tree is grown until the subset at every leaf node has entropy 0 (all observations are from the same class).

Another possible criterion choice is the Gini Index,

$$G = 1 - \sum_i p_i^2.$$

At each node, the split with minimal Gini Index should be chosen. As with the entropy criterion, the classification tree is grown until the subset at every leaf node has Gini Index 0.

Finally, classification error may be used as a criterion,

$$CE = 1 - \max_i p_i.$$

Once again, the split with minimal classification error should be chosen, and the classification tree is grown until the subset at every leaf node has classification error 0.

Our random forest modeling uses the `randomForest` package in `R`, which only supports the use of the Gini Index as a splitting criterion. However, the choice of impurity measure has been shown to have little effect on the performance of decision tree algorithms, due to their consistency.

4.4.3.2 Sampling Ratios

In this data set, the composition of binary outcomes pertaining to whether or not a hotel has defaulted on an account is approximately 1:5 for both the training and testing data sets (around five times more hotels have favorable default outcomes than unfavorable outcomes). The composition of binary outcomes pertaining to a hotel's level of activity is approximately 1:4 for both the training and testing data sets (around four times more hotels have unfavorable activity outcomes than favorable outcomes). While this level of imbalance is not extreme, we built models both with no sampling ratio adjustment and with down sampling (where each class is equally represented for training the models). In the latter set of models, the number of hotel observations sampled from each class (0,1) for building each model is equal to the number of hotel observations belonging to the smallest class. The models built using down sampling performed slightly better (improvement in KS on the order of 0.01), and these are the models for which we present results.

4.4.3.3 Number of Trees and Dimensions Considered for Splitting

All random forest models in this analysis have 500 trees. We find this number of trees is sufficiently large to achieve good quality predictions, while still computationally inexpensive. The number of dimensions to consider for splitting at each node is selected based on minimizing the out of bag (OOB) error estimate with the `tuneRF`

function in R. Out of bag samples are those left out of the bootstrap sample used in the construction of a given tree. To calculate the OOB error estimate, first classify each observation left out in constructing the kth tree using the kth tree. Then record the class with a majority of the votes every time a given observation is out of bag. Then the proportion of times the recorded class is not equal to the true class averaged over all observations is the OOB error estimate. Breiman (2001) asserts that this estimate has proven to be unbiased in many tests.

4.4.4 Results

Each of the four random forest models described in the previous section is built using the methodology outlined above. In addition, we fit two logistic regression models in order to combine the bureau and social media information. For the default model, we find $\hat{\beta}_0 = 1.159$ (p-value: 0.005), $\hat{\beta}_1 = 1.246$ (p-value: 0.082), $\hat{\beta}_2 = -3.642$ (p-value: $1.3\text{x}10^{-6}$), and $\hat{\beta}_3 = -3.661$ (p-value: 0.005). Note that while the social media related coefficient is only marginally significant, the coefficient corresponding to the interaction term is highly significant. For the activity model, we find $\hat{\beta}_0 = 1.895$ (p-value: $1.4\text{x}10^{-4}$), $\hat{\beta}_1 = 0.595$ (p-value: 0.501), $\hat{\beta}_2 = -5.724$ (p-value: $2.8\text{x}10^{-11}$), and $\hat{\beta}_3 = -0.4106$ (p-value: 0.782). In this case, neither the main effect or interaction effect involving the social media class votes is significant.

Practitioners typically evaluate the performance of models according to various criteria. For binary classification models, the Kolmogorov-Smirnov statistic is often taken to be the most important single-value summary. While values vary from application to application, KS statistics for bureau default models are usually at least 0.50. Marketing models tend to be less predictive, and a KS of at least 0.40 is often considered to be a strongly performing model. While serving as a simple and easy to consume measure of the overall predictiveness of a binary classifier, practitioners should also take other measures into account, including unit-level and dollar-level tail

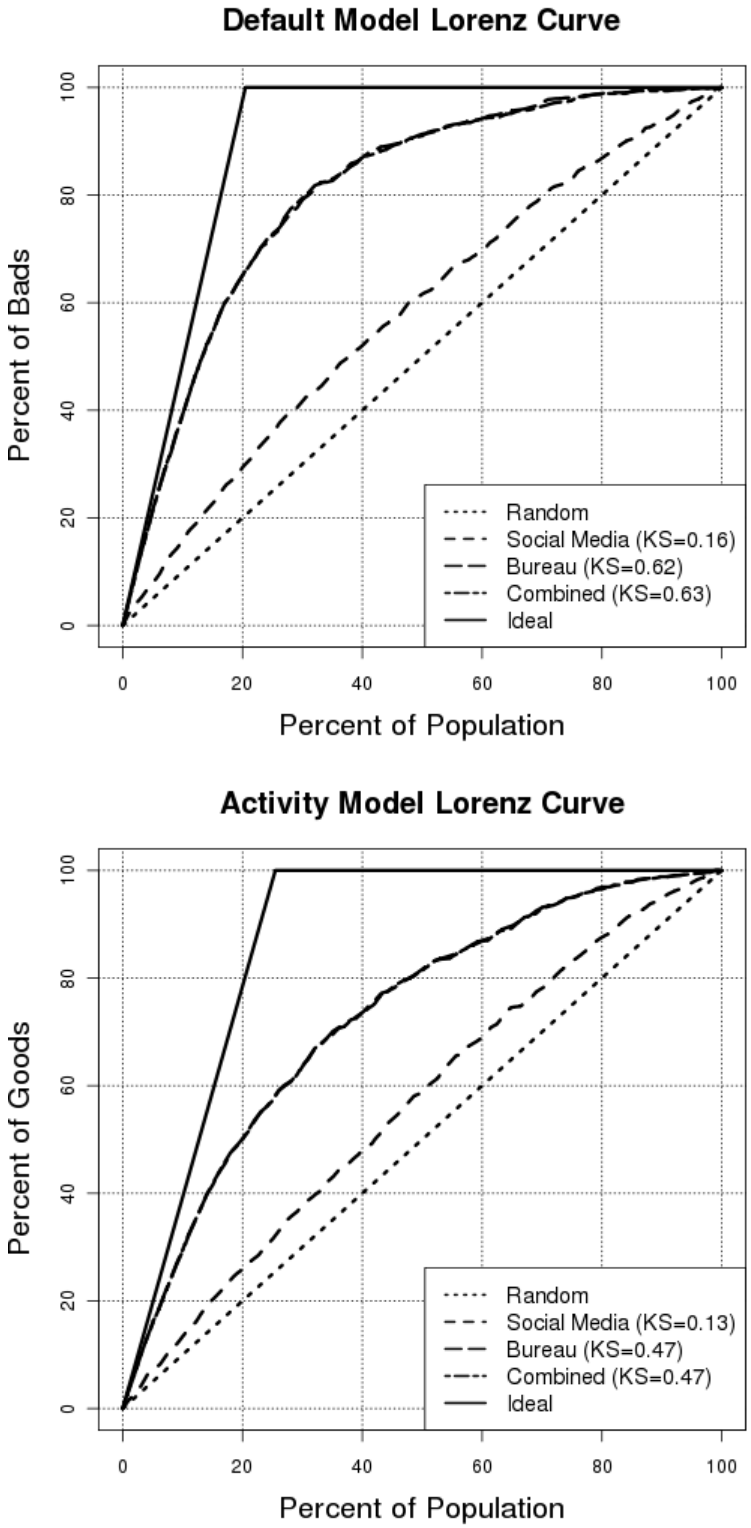


Figure 43: A comparison of the models' performances for the default dependent variable (top) and the activity dependent variable (bottom). Note that for both sets of models, the bureau only and combined curves are overlapping.

Table 10: Bad capture rates for the social media data only, traditional bureau data only, and combined default models.

Percentage of Accounts	**Social Media Data Model Capture Rates**	**Bureau Data Model Capture Rates**	**Combined Model Capture Rates**
5	7.73	21.35	21.19
10	15.95	39.18	39.10
20	29.33	65.06	65.06
30	41.83	79.17	79.49
40	52.08	86.86	87.02
50	61.54	91.35	91.03
60	69.87	94.23	94.07
70	79.97	97.36	96.47
80	87.02	98.88	98.72
90	93.59	99.68	99.36
100	100.00	100.00	100.00

Table 11: Cumulative bad dollar capture for the social media data only, traditional bureau data only, and combined default models.

Percentage of Accounts	**Social Media Data Model**	**Bureau Data Model**	**Combined Model**
5	$248,988	$197,656	$195,043
10	$686,499	$1,098,030	$1,093,064
20	$1,179,321	$3,203,637	$3,203,637
30	$1,634,684	$3,700,750	$3,713,817
40	$2,225,546	$3,872,316	$3,871,749
50	$2,352,068	$3,891,690	$3,894,842
60	$2,450,625	$3,939,690	$3,938,610
70	$2,535,524	$3,982,719	$3,982,619
80	$2,632,496	$3,983,277	$3,982,922
90	$3,602,837	$3,983,331	$3,983,331
100	$3,983,331	$3,983,331	$3,983,331

Table 12: Diagnostics for the default models

	Social Media	**Bureau**	**Combined**
KS	0.16	0.62	0.63
PI	20.7	76.7	76.4
Divergence	0.12	3.52	4.38
Info Value	0.14	2.72	2.80

capture rates.

Figure 43 compares the model performance on the testing set for the models built. This figure demonstrates that the logistic regression model, which incorporates both traditional bureau data and social media data, does not provide any additional lift in accuracy over the random forest model with only traditional bureau attributes. The Kolmogorov-Smirnov (KS) statistic for both classification procedures is 0.471 for the activity model (p-values: $2.2\text{x}10^{-16}$). For the default model, there is only a slight difference in KS (KS=0.621 for the bureau only model, KS=0.628 for the combined model, p-values: $2.2\text{x}10^{-16}$). However, the social media data itself does contain some predictive power, with both default and activity classifications better than random (KS=0.133 for the activity model with p-value $2.4\text{x}10^{-9}$, KS=0.158 for the default model with p-value $3.8\text{x}10^{-11}$).

Tables 10 and 13 summarize the bad and good capture rates, respectively, for the social media only, bureau data only, and combined models. For example, the bureau data only and combined default models are both able to capture approximately 65% of the bad accounts in the worst 20% scoring accounts while the social media only model is able to capture only around 29%.

Tables 11 and 14 additionally summarize the dollar capture rates. Our models were not trained specifically to capture the accounts with highest default dollars or highest spending on new accounts, although this could be done by giving certain observations in the training set more weight than others. However, it is still important

Table 13: Good capture rates for the social media data only, traditional bureau data only, and combined activity models.

Percentage of Accounts	**Social Media Data Model Capture Rates**	**Bureau Data Model Capture Rates**	**Combined Model Capture Rates**
5	7.09	15.88	15.75
10	13.40	29.38	29.45
20	25.64	50.00	50.00
30	37.63	63.47	63.72
40	48.07	73.45	73.58
50	58.76	81.70	81.44
60	69.07	86.98	86.86
70	78.03	93.11	92.98
80	87.63	96.78	96.52
90	94.91	98.84	98.84
100	100.00	100.00	100.00

Table 14: Cumulative good dollar capture for the social media data only, traditional bureau data only, and combined activity models.

Percentage of Accounts	**Social Media Data Model**	**Bureau Data Model**	**Combined Model**
5	$97,961	$336,980	$347,430
10	$206,137	$593,275	$598,099
20	$505,108	$920,022	$923,195
30	$886,692	$1,216,805	$1,170,571
40	$1,048,032	$1,402,797	$1,415,226
50	$1,271,163	$1,828,479	$1,823,099
60	$1,481,366	$2,075,082	$2,071,750
70	$1,663,311	$2,209,002	$2,212,106
80	$2,017,092	$2,336,280	$2,335,151
90	$2,206,078	$2,385,426	$2,385,440
100	$2,434,025	$2,434,025	$2,434,025

Table 15: Diagnostics for the activity models

	Social Media	**Bureau**	**Combined**
KS	0.13	0.47	0.47
PI	17.4	61.0	60.9
Divergence	0.12	1.62	2.00
Info Value	0.13	1.50	1.46

to understand the models' performance in these terms. Without considering dollar capture, whether our activity models tend to capture $1 or $1,000 new accounts, for example, is unknown. Tables 12 and 15 provide additional summaries for each model, including predictiveness index (PI), divergence, and information value.

4.4.4.1 Variable Importance

Tables 16 and 17 list the top five most influential attributes in each of the random forest models. We measure importance through the mean decrease in Gini Index. A higher decrease in Gini Index means that a particular predictor variable plays a greater role in partitioning the data into the defined classes. For the social media only model, the rating attributes are of much greater importance than the review title attributes, with five to ten times larger mean decreases in Gini Index. The most

Table 16: Top five most influential attributes in the social media only and bureau only default models.

Attribute	**Mean Decrease Gini**
Average neighborhood rating through Sept	75.9
Average comfort rating through Sept	73.5
Overall average rating through Sept	72.8
Average service rating through Sept	72.8
Average condition rating through Sept	72.4
Total non-fi past due amount in last 6 months	50.7
Worst non-fi payment status in last 3 months	45.4
Non-fi accounts 2+ cycles past due or charge-off in last 3 months	44.1
Non-fi accounts 3+ cycles past due or charge-off in last 3 months	35.8
Percentage non-fi past due to total reported balance in last 3 months	34.8

Table 17: Top five most influential attributes in the social media only and bureau only activity models.

Attribute	**Mean Decrease Gini**
Average service rating through Sept	90.2
Average neighborhood rating through Sept	88.4
Average condition rating through Sept	85.9
Overall average rating through Sept	84.5
Average cleanliness rating through Sept	84.0
Highest industry balance in last 3 months	53.4
Highest non-fi balance in last 3 months	41.9
Highest industry balance in last 12 months	40.7
Highest industry balance in last 24 months	37.8
Highest non-fi balance in last 12 months	33.0

influential bureau attributes all pertain to non-financial trade lines, such as past due amounts or account balances.

4.5 Discussion and Conclusions

In this chapter, we have discussed traditional credit modeling methodologies as well as new trends in decisioning based on alternative sources of data. We have also described the random forest methodology and its advantages in modeling credit data. In this context, we have examined the effectiveness of using online business review data in the commercial space specifically to make decisions whether to lend and/or target for marketing purposes. Using modeling attributes based on online reviews for over 43,000 U.S.-based hotels and 400 traditional credit bureau attributes, we have built two sets of default and activity models. Finally, we have evaluated using the review data both as a stand alone decision making tool and as an addition to models built with only bureau attributes.

Two aspects of the modeling results are particularly interesting. The first is the fact that useful models can be built using only the hotel online review data. For the default and activity models, the KS statistics were 0.158 and 0.133, respectively.

Each of these values corresponds to a highly significant p-value ($3.8\text{x}10^{-11}$, $2.4\text{x}10^{-9}$). Furthermore, the Lorenz curves and capture rates demonstrate that these models allow lenders to somewhat target certain hotels based on their default likelihood and level of activity. For example, the default model built only on social media attributes was able to capture 29.3% of all bad accounts in the bottom-scoring 20% of the distribution, allowing lenders to avoid these more risky prospects. While not nearly as strong as the corresponding values for the bureau models, if a practitioner had nothing available besides such models on which to decision, they certainly do provide some predictive value.

The other interesting comparison we wish to draw stems from adding the hotel review attributes into an existing bureau model, in order to determine whether any performance lift is to be found by doing so. For the default and activity bureau models, the KS statistics were 0.621 and 0.471, respectively (p-values: $2.2\text{x}10^{-16}$, $2.2\text{x}10^{-16}$). When expanding the modeling attribute set to include the hotel review data, the KS statistics were 0.628 and 0.471 (p-values: $2.2\text{x}10^{-16}$, $2.2\text{x}10^{-16}$). As an example of tail capture, we consider the default dollars captured in the bottom 10% of the distribution, which were \$1,098,030 and \$1,093,064 for the respective models, comprising percentages of overall bad dollars of 27.6% and 27.4%, respectively. Apart from a slight fluctuation due to noise, the model enriched with the social media attributes performs no better than the existing bureau model. This stems largely from the fact that, while some predictive signal exists in the social media information, the magnitude of the signal pales in comparison to that of the bureau data with respect to the dependent variables we are trying to model.

APPENDIX A

EXPRESSION 20 PROOF

A.0.1 Derivation of expression (20)

The scale-mixing detail coefficients of a 2-D fBm are

$$d_{(j,j+s);\mathbf{k}} = 2^{j+\frac{s}{2}} \int B_H(\mathbf{u}) \psi \left(2^j u_1 - k_1, 2^{j+s} u_2 - k_2\right) d\mathbf{u},$$

where ψ stands for ψ^d in (8). These coefficients are random variables with zero mean and variance (Heneghan et al., 1996)

$$\begin{aligned} \mathbb{E}\left[d^2_{(j,j+s);\mathbf{k}}\right] = 2^{2j+s} \int & \psi \left(2^j u_1 - k_1, 2^{j+s} u_2 - k_2\right) \\ & \times \psi \left(2^j v_1 - k_1, 2^{j+s} v_2 - k_2\right) \mathbb{E}\left[B_H(\mathbf{u}) B_H(\mathbf{v})\right] d\mathbf{u}\, d\mathbf{v}. \end{aligned} \tag{24}$$

Since

$$\mathbb{E}\left[B_H(\mathbf{u}) B_H(\mathbf{v})\right] = \frac{\sigma_H^2}{2} \left(|\mathbf{u}|^{2H} + |\mathbf{v}|^{2H} - |\mathbf{u} - \mathbf{v}|^{2H}\right),$$

and

$$\int \psi \left(2^j u_1 - k_1, 2^{j+s} u_2 - k_2\right) d\mathbf{u} = \int \psi \left(2^j v_1 - k_1, 2^{j+s} v_2 - k_2\right) d\mathbf{v} = 0,$$

it can be easily seen that (24) becomes

$$\begin{aligned} \mathbb{E}\left[d^2_{(j,j+s);\mathbf{k}}\right] = -\frac{\sigma_H^2}{2}\, 2^{2j+s} \int \int & \psi \left(2^j u_1 - k_1, 2^{j+s} u_2 - k_2\right) \\ & \times \psi \left(2^j v_1 - k_1, 2^{j+s} v_2 - k_2\right) |\mathbf{u} - \mathbf{v}|^{2H} d\mathbf{u}\, d\mathbf{v}. \end{aligned}$$

Next, define substitutions

$$\begin{aligned} \mathbf{p} &= (p_1, p_2) = \left(2^j (u_1 - v_1), 2^{j+s} (u_2 - v_2)\right), \\ \mathbf{q} &= (q_1, q_2) = \left(2^j v_1 - k_1, 2^{j+s} v_2 - k_2\right) \end{aligned}$$

Then, if $\mathbf{p}_s \equiv (p_1, 2^{-s}p_2)$,

$$\begin{aligned}
\mathbb{E}\left[d^2_{(j,j+s);\mathbf{k}}\right] &= -\frac{\sigma_H^2}{2}\, 2^{2j+s} \int\int \psi(\mathbf{p}+\mathbf{q})\ \psi(\mathbf{q})\ 2^{-2jH}\ |\mathbf{p}_s|^{2H}\ 2^{-4j-2s}\ d\mathbf{p}\ d\mathbf{q} \\
&= -\frac{\sigma_H^2}{2}\, 2^{-j(2H+2)} \int\int \psi(\mathbf{p}+\mathbf{q})\ \psi(\mathbf{q})\ |\mathbf{p}_s|^{2H} 2^{-s}\ d\mathbf{p}\ d\mathbf{q} \\
&= 2^{-j(2H+2)}\ V_{\psi,s}(H),
\end{aligned}$$

where

$$V_{\psi,s}(H) = -\frac{\sigma_H^2}{2} \int\int \psi(\mathbf{p}+\mathbf{q})\ \psi(\mathbf{q})\ |\mathbf{p}_s|^{2H} 2^{-s}\ d\mathbf{p}\ d\mathbf{q}, \tag{25}$$

is an integral depending on ψ, H, and s but not on the scale j.

Bibliography

Aberle DR, A. A., CD, B., WC, B., JD, C., RM, F., IF, G., C, G., PM, M., and JD, S. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409.

Afreixo, V., Ferreira, P. J., and Santos, D. (2004). Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*, 14(6):523–530.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Reports*, 1(5):543–556.

Armour, S. (2014). Borrowers hit social-media hurdles. *Wall Street Journal Online*.

Arneodo, A., Vaillant, C., Audit, B., Argoul, F., dAubenton Carafa, Y., and Thermes, C. (2011). Multi-scale coding of genomic information: From DNA sequence to genome structure and function. *Physics Reports*, 498(2):45–188.

Bai, F. L., Liu, Y. Z., and Wang, T. M. (2007). A representation of DNA primary sequences by random walk. *Mathematical Biosciences*, 209(1):282–291.

Barnard, G. (1956). Discussion of Hurst. *Proc. Inst. Civ. Eng.*, 5:552–553.

Beran, J. (1994). *Statistical Methods for Long Memory Processes*. Chapman & Hall, London.

Berzina, D., Nakazawa-Miklasevica, M., Zestkova, J., Aksenoka, K., Irmejs, A., Gardovskis, A., Kalniete, D., Gardovskis, J., and Miklasevics, E. (2013). BCRA1/2 mutation screening in high-risk breast/ovarian cancer families and sporadic cancer patient surveilling for hidden high-risk families. *BMC Medical Genetics*, 14:61.

Boekhorst, R. T., Abnizova, I., and Nehaniv, C. (2008). Discriminating coding, non-coding and regulatory regions using rescaled range and detrended fluctuation analysis. *BioSystems*, 91(1):183–194.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brown, V. R. and Vaughn, E. D. (2011). The writing on the (Facebook) wall: The use of social networking sites in hiring decisions. *Journal of Business and Psychology*.

Buldyrev, S., Goldberger, A., Havlin, S., Mantegna, R., Matsa, M., Peng, C.-K., Simons, M., and Stanley, H. (1995). Long-range correlation properties of coding and noncoding DNA sequences: Genbank analysis. *Physical Review E*, 51(5):5084.

Carney, P., Miglioretti, D., Yankaskas, B., Kerlikowske, K., Rosenberg, R., Rutter, C., Geller, B., Abraham, L., Taplin, S., Dignan, M., Cutter, G., and Ballard-Barbash, R. (2003). Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Ann Intern Med*, 138(3):168–175.

Cattani, C., Scalia, M., and Mattioli, G. (2006). Entropy distribution and information content in DNA sequences. In *International Conference on Potential Theory and Complex Analysis*, pages 8–11.

Chan, H. P., Goi, K., Galhotra, S., Vyborny, C. J., MacMahon, H., and Jokich, P. M. (1987). Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography. *Medical Physics*, 14(4):538–548.

Cheng, H. D., Cai, X., Chen, X., Hu, L., and Lou, X. (2003). Computer-aided detection and classification of microcalcifications in mammograms: A survey. *Pattern Recognition*, 36(12):2967–2991.

Ching-Chiang, Y., Fengyi, L., and Chih-Yu, H. (2012). A hybrid KMV model, random forests and rough set theory approach for credit rating. *Knowledge-Based Systems*, 33:166–172.

Commission, U. F. T. (1998). Facts for consumers.

Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Deutsch, M. and Long, M. (1999). Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res*, 27(15):3219–3228.

Dixon, P. and Gelman, R. (2014). The scoring of America: How secret consumer scores threaten your privacy and your future. *worldprivacyforum.org*.

Elder, D. (2000). Split gene origin and periodic introns. *Journal of Theoretical Biology*, 207(1):455–472.

Embrechts, P. and Maejima, M. (2002). *Selfsimilar Processes*. The London Mathematical Society (Cambridge), Academic Press.

Feller, W. (1951). The asymptotic distribution of the range of sums of independent random variables. *Ann. Math. Statist.*, 22:427–432.

Finlay, S. (2010). *Credit Scoring, Response Modeling, and Insurance Rating: A Practical Guide to Forecasting Consumer Behavior*. Palgrave Macmillan, Basingstoke, UK.

Flandrin, F. (1992). Wavelet analysis and synthesis of fractional Brownian motion. *IEEE Transactions on Information Theory*, 38(2):910–917.

Freer, T. W. and Ulissey, M. J. (2001). Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center. *Radiology*, 220(3):781–786.

Goel, S. and Goldstein, D. G. (2013). Predicting individual behavior with social networks. *Marketing Science*, 33(1):82–93.

Goldstraw, P. and Crowley, J. (2007). The IASLC lung cancer staging project: Proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM classification of malignant tumors. *Journal of Thoracic Oncology*, 2(8):706–714.

Guan, W., Zhou, M., Hampton, C. Y., Benigno, B. B., Walker, L. D., Gray, A., and McDonald, J. (2009). Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics*, 10:259.

Haimovich, A. D., Byrne, B., Ramaswamy, R., and Welsh, W. J. (2006). Wavelet analysis of DNA walks. *Journal of Computational Biology*, 13(7):1289–1298.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, USA.

Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, W. P., Moore, R., Chang, K., and Kumaran, S. M. (1998). Current status of the digital database for screening mammography. *Proceedings of the Fourth International Workshop on Digital Mammography*, pages 457–460.

Heath, M., Bowyer, K., Kopans, D., Moore, R., and Kegelmeyer, W. P. (2001). The digital database for screening mammography. *Proceedings of the Fifth International Workshop on Digital Mammography*, pages 212–218.

Heneghan, C., Lown, S., and Teich, M. (1996). Two dimensional fractional Brownian motion: Wavelet analysis and synthesis. *Image analysis and interpretation, proceedings of the IEEE Southwest Symposium*, pages 213–217.

Hill, S. F. P. and Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):256–276.

Hostalkova, E. and Prochazka, A. (2007). Complex wavelet transform in biomedical image denoising. *Proceedings of 15th Annual Conference Technical Computing Prague*, pages 1–8.

Houssami, N., Irwig, L., and Ciatto, S. (2006). Radiological surveillance of interval breast cancers in screening programmes. *The Lancet Oncology*, 7(3):259–265.

Ieviņa, N., Chipens, G., and Kalvinsh, I. (2006). Internal regularity and quantization of gene parameters. *Acta Universitatis Latviensis*, 710(1):139–153.

Jung, Y. Y., Park, Y., Jones, D., Ziegler, T., and Vidakovic, B. (2010). Self-similarity in NMR spectra: An application in assessing the level of cysteine. *Journal of Data Science*, 8(1):1 – 19.

Klik, M. A. J., v Rikxoort, E. M., Peters, J., Gietema, H. A., Prokop, M., and v Ginneken, B. (2006). Improved classification of pulmonary nodules by automated detection of benign sub pleural lymph nodes. *Conference: 2006 3rd IEEE International Symposium on Biomedical Imaging*, pages 494–497.

Kruppa, J., Schwarz, A., Arminger, G., and Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13):5125–5131.

Lamperti, J. (1972). Semi-stable stochastic processes. *Amer. Math. Soc. Trans.*, 104:62–78.

Lin, M., Prabhala, N. R., and Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, 59(1):17–35.

Lohr, S. (2015). Banking start-ups adopt new tools for lending. *New York Times Online*.

Magee, C. (2015). Trillion-dollar alternative lending industry is a VC gold mine. *http://techcrunch.com/2015/03/05/trillion-dollar-alternative-lending-industry-is-a-vc-gold-mine*.

Mallat, S. (1997). *A Wavelet Tour of Signal Processing*. AP Professional, London.

Mandelbrot, B. (1975). *Les objects fractals: forme, hasard et dimension*. Flammarion, Paris.

Mandelbrot, B. and J. W. Van Ness, J. W. (1968). Fractional brownian motions, fractional noises and applications. *SIAM rev.*, 10(4):422–437.

Mandelbrot, B. and Wallis, J. R. (1969). Some long-run properties of geophysical records. *Water Resources Res. 5*, 5:321–340.

Mays, E. (2001). *Handbook of Credit Scoring*. Glenlake Publishing, Chicago, USA.

Meziane, M. and et al. (2012). A comparison of four versions of a computer-aided detection system for pulmonary nodules on chest radiographs. *Journal of thoracic imaging*, 27(1):58–64.

National Cancer Institute (2014). Mammograms fact sheet. *http://www.cancer.gov/cancertopics/types/breast/mammograms-fact-sheet*.

Nicolis, O., Jeon, S., and Vidakovic, B. (2012). Mammogram diagnostics via 2-D complex wavelet-based self-similarity measures. *Quaderni del Dipartimento di Ingegneria dell'informazione e metodi matematici.*

Nicolis, O., Ramírez-Cobo, P., and Vidakovic, B. (2010). 2-D wavelet-based spectra with applications. To appear in *Computational Statistics and Data Analysis*.

Oda, S. and et all. (2010). Detection of small pulmonary nodules on chest radiographs: Efficacy of dual-energy subtraction technique using flat-panel detector chest radiography. *Clinical Radiology*, 65(8):609–615.

Oppenheim, G. and Taqqu, M. S. (2003). Fractional brownian motion and long-range dependence. In *Theory and Applications of Long-range Dependence*, pages 5–38. P. Doukhan, G. Oppenheim & M. S. Taqqu, eds.,Birkhauser.

Paxia, S., Rudra, A., Zhou, Y., and Mishra, B. (2002). A random walk down the genomes: DNA evolution in valis. *Computer*, 35(7):73–79.

Peng, C.-K., Buldyrev, S., Goldberger, A., Havlin, S., Sciortino, F., Simons, M., and Stanley, H. (1992). Long-range correlations in nucleotide sequences. *Nature*, 356(1):168–170.

Pinho, A. J., Neves, A. J., Afreixo, V., Bastos, C. A., and Ferreira, P. J. S. (2006). A three-state model for DNA protein-coding regions. *Biomedical Engineering, IEEE Transactions on*, 53(11):2148–2155.

Ramírez-Cobo, P., Lee, K. S., Molini, A., Porporato, A., Katul, G., and Vidakovic, B. (2011). A wavelet-based spectral method for extracting self-similarity measures in time-varying two-dimensional rainfall maps. *Journal of Time Series Analysis*, 32(4):351–363.

Roth, P. L., Bobko, P., Van Iddekinge, C. H., and Thatcher, J. B. (2013). Social media in employee-selection-related-decisions: A research agenda for uncharted territory. *Journal of Management*.

Rusli, E. M. (2013). Bad credit? Start tweeting. *Wall Street Journal Online*.

Sakharkar, M. K. and Chow, V. T. (2004). Distributions of exons and introns in the human genome. *In Silico Biol*, 4(4):387–393.

Samorodnitsky, G. and Taqqu, M. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall, New York.

Sharma, D. (2012). Improving the art, craft, and science of economic credit risk scorecards using random forests: Why credit scorers and economists should use random forests. *Academy of Banking Studies Journal*, 11(1):93–116.

Shi, B., Moloney, K. P., Pan, Y., Leonard, V. K., Vidakovic, B., Jacko, J., and Sainfort, F. (2006). Classification of high frequency pupillary responses using Schur monotone descriptors in multiscale domains. *Journal of Statistical Computation and Simulation*, 76:431–446.

Stoffer, D. S., Tyler, D. E., and Wendt, D. A. (2000). The spectral envelope and its applications. *Statistical Science*, pages 224–253.

The Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, 443(26):931–949.

Tibshirani, R. (2007). A comparison of fold-change and the t-statistic for microarray data analysis. *Analysis*, pages 1–17.

Veitch, D. and Abry, P. (1999). A wavelet-based joint estimator of the parameters of long-range dependence. *IEEE Transactions on Information Theory*, 45:878–897.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. John Wiley & Sons, Inc.

Vinogradov, A. (2006). 'Genomic design' model and multicellular complexity: Golden middle. *Nucleic Acids Research*, 34(20):5906–5914.

Voss, R. F. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters*, 68(25):3805.

Wei, Y., Yildirim, P., Van den Bulte, C., and Dellarocas, C. (2014). Credit scoring with social network data. *Social Science Research Network*.

Williams, T. I., Toups, K. L., Saggese, D. A., Kalli, K. R., Cliby, W. A., and Muddiman, D. C. (2007). Epithelial ovarian cancer: Disease etiology, treatment, detection, and investigational gene, metabolite, and protein biomarkers. *Journal of Proteome Research*, 6(8):2936–2962.

Yin, C. and Yau, S. S.-T. (2007). Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *Journal of Theoretical Biology*, 247(4):687–694.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J.-Q., and Tian, D. (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics*, 10(1):47.

VITA

Tonya Woods is a Ph.D. candidate in the School of Industrial & Systems Engineering (ISyE) at the Georgia Institute of Technology with a specialization in statistics. She joined the Ph.D. program in 2012 and was taken under the advisement of Dr. Brani Vidakovic shortly thereafter. Her research interests include high dimensional data classification techniques, wavelets, machine learning, and data science. She has been working with Equifax since May 2014 as a statistical consultant and plans to continue this work after graduation. Before beginning the Ph.D. program, she earned a Bachelor of Science degree in Mathematics as well as a Bachelor of Arts degree in Dance Performance from the University of Alabama in 2010, and she earned a Masters degree in Statistics from ISyE at Georgia Tech in 2012.